BAYESIAN NETWORK MODELING AND INFERENCE IN PLANT GENE NETWORKS AND ANALYSIS OF SEQUENCING AND IMAGING DATA

A Dissertation

by

PRIYADHARSHINI SUNDARARAJAN VENKATASUBRAMANI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Aniruddha Datta
Co-Chair of Committee,	Krishna R. Narayanan
Committee Members,	Shankar P. Bhattacharyya
	Byung-Jun Yoon
	Anirban Bhattacharya
Head of Department,	Miroslav Begovic

August 2017

Major Subject: Electrical Engineering

ABSTRACT

Scientific and technological advancements over the years have made curing, preventing or managing all diseases, a goal that seems to be within reach. The approach to manipulating biological systems is multifaceted. This dissertation focuses on two problems that pose fundamental challenges in developing methods to control biological systems: the first is to model complex interactions in biological systems; the second is faithful representation and analysis of biological data obtained from scientific equipments.

The first part of this dissertation is a discussion on modeling and inference in gene networks, and Bayesian inference. Then we describe the application of Bayesian network modeling to represent interactions among genes, and integrating gene expression data in order to identify potential points of intervention in the gene network. We conclude with a summary of evolving directions for modeling gene interactions.

The second topic this dissertation focuses on is taming biological data to obtain actionable insights. We introduce the challenges in representation and analysis of high throughput sequencing data and proceeds to describe the analysis of imaging data in the dynamic environment of cancer cells. Then we discuss tackling the problem of analyzing high throughput RNA sequencing data in order to pinpoint genes that exhibit different behaviors under monitored experimental conditions. Then we address the interesting problem of deciphering and quantifying gene-level activity from epifluorescent imaging data.

ACKNOWLEDGEMENTS

I am very grateful to my advisor Dr. Aniruddha Datta for his guidance, support, encouragement, and for giving me the freedom to pursue certain topics that piqued my interest. I would like to thank my co-advisor Dr. Krishna Narayanan for his valuable advice and the many thought-provoking questions, and my committee members Dr. Byung-Jun Yoon, Dr. Shankar Bhattacharyya and Dr. Anirban Bhattacharya for the support and inspirational comments.

I would like to thank Dr. Charlie Johnson and Dr. Michael Bittner for eye-opening discussions on plant and animal biology. I am grateful to many other Professors with whom I had the opportunity to take interesting courses that I thoroughly enjoyed. I am also grateful for the opportunity to attend numerous stimulating seminars and lecture series, especially the CESG Fishbowl and ISS seminars. I am also grateful to the ECEN Department, CBGSE and NSF for funding my research through various grants.

I would like to thank my past and present labmates Sriram, Anwoy, Bibhu, Osama, and Ashish for advice and friendship, and my friends Bindu, Prerna, Udit, and a bunch of great Aggies, for providing me with various combinations of food, shelter, and companionship, over the years. I would also like to thank my peers Shishir, Lakshmi Narasimhan and Jayavel for inspiration and guidance. Last, but by no means the least, I would like to thank my parents Sasikala and Venkatasubramani, and my brother Balaji, for their love and support, and for encouraging me to pursue my dreams.

CONTRIBUTORS AND FUNDING SOURCES

The work in this dissertation was carried out at the Department of Electrical and Computer Engineering and the TEES-Agrilife Centre for Bioinformatics and Genomics Systems Engineering (CBGSE). The major contributors to the material presented in this thesis include my committee Chair Dr. Aniruddha Datta and my committee Co-Chair Dr. Krishna Narayanan. Other major contributors are Dr. Chao Sima, Dr. Jianping Hua, Dr. Michael Bittner and Milana Cypert.

The plant sequencing data analyzed in Chapter 3 was generated by CBGSE, while the experimental setup and sample details were provided by Dr. Cecilia Tamborindeguy and Ordom Huot from the Department of Entomology.

The research presented in this dissertation was funded in part by the National Science Foundation under grant [ECCS-1404314] and in part by the AgriLife-TEES Center for Bioinformatic and Genomic Systems Engineering (CBGSE).

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

"Prediction is very difficult, especially about the future."

– Niels Bohr

Decision making is an integral part of our everyday lives. We make these decisions based on what can be inferred from modeling our observations. The modeling is carried out in a way that enables capturing significant characteristics of the data in order to predict future events.

Existing knowledge can sometimes be useful in constructing the model, but many a time, the underlying processes that generate the data are not understood well enough that an accurate model can be designed. In such cases, we try to create models that are good approximations of the processes through which the data was generated. For example, trying to predict the weather for the next few months, knowing the temperature, humidity, wind patterns, etc. of the past. Since we cannot see into the future and know how tomorrow's weather would be, we rely on a model of the climatic conditions in order to be able to perform the predictions.

To begin modeling the data, we use a reasonably flexible model specified by a set of parameters, and then attempt to find settings of these parameters that explain the observed data in the best possible manner. The methodology by which we fit model parameters to observed data is called learning the model. Our belief is that once we have a satisfactory model that explains observed data well, we can confidently use the model to predict future observations.

Since these models are approximations of the real processes, there will be characteristics of the data that we do not capture in these models, which are considered noise. When it comes to tuning the parameters of our model, it often becomes diffi-

cult to know which settings of the parameters capture nuances of the data that are relevant for predicting future observations, and which settings capture noise. It may happen a very complex model's parameters fit data exactly, but since this model captured noise characteristics, the predictions obtained using this model will not be optimal. On the other hand, a very simple model may not capture the patterns in the data and therefore will also not provide optimal predictions. There are several studies that address this trade-off between model complexity and generalization, more on this to be discussed later.

The framework of Bayesian inference can be used to fomalize the above concepts of of using parametric models to fit data. Let y denote the data set such that $y = y_1, y_2, ..., y_k, ..., y_K$. The data can be defined by a model that has the parameters $\Theta = \theta_1, \theta_2, ..., \theta_J$ such that Θ define a probability distribution $p(y|\theta)$.

There are several ways of finding the parameters to learn the model. The maximum likelihood approach involves finding parameters θ^* such that the likelihood of θ, or the probability of observing the data under the model is maximal.

$$\theta^* = \arg\max_{\theta} p(y|\theta) \tag{1.1}$$

The model may also include other hidden variables that have not been observed, but have an effect on the observed data through the parameters. The probability distribution of the data can then be written as:

$$p(y|\theta) = \sum_{x} p(x|\theta)p(y|x,\theta) \tag{1.2}$$

where we denote the hidden variables by x, and we sum over all possible hidden states.

Then, the *posterior* distribution over the hidden variables, can be estimated using

Bayes' rule as:

$$p(x|y,\theta) = \frac{p(x|\theta)p(y|x,\theta)}{p(y|\theta)} \tag{1.3}$$

Here, $p(x|\theta)$ is called the *prior* probability of the hidden variables. The prior is usually set in such a way as to to reflect the distribution of data that the modeler expects. In section 1.1, we provide a review of graphical models and how they can be used to visualize independence relationships between different variables in the model. In section 1.2, we briefly introduce literature on statistical modeling of biological systems.

1.1 Probabilistic Graphical Models

Statistical modeling typically involves multiple random variables that interact with each other. Graphical models provide a way to conveniently represent relationships between these variables. A number of statistical models can be naturally expressed using probabilistic graphical models. A probabilistic graphical model represents a family of probability distributions on variables in the model. Each variable is represented by a node, and an edge between two nodes in the model indicates a connection between them. The pattern of edges between the different nodes comprise the structure of the model.

A class of graphical models called *directed graphical models* , or *Directed Acyclic Graphs* (DAGs), contain directed edges between the variables. The directed graphical model is called an *acyclic* graph since no directed *paths* exist such that the same variable is encountered more than once. There is another class of models called Undirected graphical models, but these are not discussed here.

Using DAGs, we can conveniently express the conditional independence relationships between variables. A random variable x is conditionally independent of y,

given z if $p(x, y|z)$ can be written $p(x|z)p(y|z)$. Using the general idea of conditional independence, the probability distribution over the nodes in a DAG can be written as:

$$p(z) = \Pi_{k=1}^{K} p(z_k|z_{pa(k)}) \tag{1.4}$$

where $z_{pa(k)}$ stand for the *parent* nodes of node k in the graph. If there is a directed edge from x to y, x is called the parent of y. The conditional independence relationships in a probabilistic graphical model can be exploited to design efficient message-passing algorithms in order to perform inference. We provide more discussion on inference algorithms in chapter 2.

1.2 Bayesian Network Modeling of Biological Systems

Genome-scale data such as gene expression data and protein expression data pose challenges since the raw data are often difficult to comprehend directly. DNA hybridization arrays measure transcription levels within the cell at a particular time for hundreds of genes. A big challenge in computational biology is uncovering interactions between genes /proteins using such measurements. Friedman et. al proposed a Bayesian network framework for analyzing interactions between genes using gene expression measurements [12]. In chapter 2, we provide a discussion on exploiting this framework to model interaction between genes in plant and pathogen systems and using message passing algorithms for approximate inference in the model.

1.3 Analysis of RNA Sequencing Data

Since the Human Genome Project published the first map of the human genome in 2004, there has been a rise in the number of new technologies that enable the study of living organisms. Sophisticated techniques such as next-generation sequencing, high-

resolution imaging, Reverse Transcription Polymerase Chain Reaction (RT-PCR), high-throughput screening, etc., generate huge amounts of raw data that can reveal valuable information about living organisms which can be used to develop effective predictive and precision medicine approaches.

One of the challenges in controlling biological system is taming and understanding the data from advanced scientific instruments. A number of research teams across the world are involved in developing mathematical methods and innovative algorithms so as to exploit the data from these sophisticated techniques more effectively.

RNA-Seq uses next generation sequencing technology to detect and measure the quantity of RNA (Ribonucleic acid) in a biological sample at any given time. The RNA-Seq platform takes chemically processed RNA as an input and outputs digital files with genetic information. Importantly, the RNA of the entire organism is not produced as one long stretch, instead, the sequencing platform produces a large set of short fragments, each containing a small fraction of genetic information. These fragments are called 'sequencing reads'. These 'raw' sequencing reads are then assembled together to reconstruct the entire *transcriptome*. In order to perform this reconstruction, two approaches are currently used: *De novo*, and *Genome guided*. The *De novo* approach is typically used when the genome to be reconstructed in unknown, or has been substantially altered, and does not require a reference genome for reconstruction. The genome guided approach uses a reference genome. Alignment algorithms that use the genome guided approach proceed in two steps: first, align short portions of the read to the reference genome and then, use dynamic programming to find an optimal alignment of the reads. For studying cellular changes in response to external stimuli, gene expression is quantified from the RNA sequence data. The number of reads that mapped to each locus during the transcriptome assembly is counted to obtain the gene expression.

In chapter 3, we provide a discussion on using RNA-Seq technology to analyze the transcriptomes of wild tomato (*Solanum lycopersicum*) plants affected by different types of stresses induced by drought, psyllid herbivory, and pathogens, individually and in combination.

1.4 Analysis of Epifluorescent Imaging Data

Image data of cellular micro environments obtained from high precision microscopes have been very helpful in understanding the mechanisms of biological systems.

Metformin, a widely used anti-diabetic drug, has recently been associated with inhibition of cell proliferation and induction of cell death in multiple cancers. The experimental evidence for the effect of metformin on induction of cell death in cancer cell lines is sparse. Besides, the mechanism and molecular basis of the action of metformin on cancer cell lines is not clearly understood.

In chapter 4, we provide a discussion on our experiments to study the effect of metformin on induction of cell death. We performed high content epifluorescent imaging analysis to quantify the amount of dead and live cells on treatment with metformin individually and in combination with chemotherapy drugs in two ovarian cancer cell lines.

2. A BAYESIAN NETWORK BASED APPROACH TO SELECTION OF INTERVENTION POINTS IN THE MAPK PLANT DEFENSE RESPONSE PATHWAY*

An important problem in computational biology is the identification of potential points of intervention that can lead to modified network behavior in a genetic regulatory network. We consider the problem of deducing the effect of individual genes on the behavior of the network in a statistical framework. In this chapter, we make use of biological information from the literature to develop a Bayesian network and introduce a method to estimate parameters of this network using data relevant to the biological phenomena under study. Then, we give a novel approach to select significant nodes in the network using a decision theoretic approach. The proposed method is applied to the analysis of the Mitogen Activated Protein Kinase (MAPK) pathway in plant defense response to pathogens. Results from applying the method to experimental data show that the proposed approach is effective in selecting genes that play crucial roles in the biological phenomenon being studied.

2.1 Introduction

The last decade has seen tremendous improvements in genetic studies and understanding of gene-level interactions in various organisms. This has been made possible as a result of high throughput data generated from microarray and sequencing technologies combined with computational modeling and simulation of interaction between various biological components in an organism. We focus on the analysis of

*Parts of this section are reprinted with permission from "A Bayesian Network-Based Approach to Selection of Intervention Points in the Mitogen-Activated Protein Kinase Plant Defense Response Pathway" by Venkat Priya S., Narayanan Krishna R., and Datta Aniruddha, 2017. *Journal of Computational Biology*, volume 24, no. 4, pages 327-339, ©2017 JCB. doi:10.1089/cmb.2016.0089.

plant-pathogen interaction in this study.

The world's growing population has made food security a global concern. One of the key factors that impact food security is the loss of crops due to diseases caused by plant pathogens [1]. Many plant-associated microbes are parasitic organisms that impair plant growth and reproduction. Plants possess inherent immune receptors that detect the presence of microbial pathogens and trigger defense responses against a multitude of harmful pathogens. But, due to adaptive evolution and the fight for survival, pathogens have developed various strategies to invade host plant tissue undetected and cause infections that render food crops unfit for consumption. A number of plant biological studies have implicated the Mitogen Activated Protein Kinase (MAPK) cascade in plant cell signaling as the point of convergence of various stress stimuli [2]. Hence, there has been a lot of research in targeting specific components of the MAPK pathway in an effort to improve disease resistance in crop plants. A mathematical model of the interactions among the components of the MAPK pathway is a critical tool in obtaining a better understanding of the nature of these interactions and advocating intervention strategies for breeding disease-resistant crops.

A number of methods have been proposed for the selection of 'important' genes from a large set of genes. The most widely mentioned among these are based on sample classification of huge volumes of genotypic and phenotypic data using numerous data mining techniques such as support vector machines (SVM) [3], regression techniques [4], random forest [5], etc. A lot of work has been presented on using clustering methods to pick significant genes ([6], [7], [8]) . Gene ranking methods use feature selection criteria to pick genes that show a lot of variation among different treatment conditions.

However, the genes selected from these methods may be irrelevant to the biological

phenomena being studied or it may even be difficult to ascribe biological connotations to the genes selected from many of these methods, since gene expression patterns alone may not convey the essence of interactions among genes.

Another popular approach for analyzing and making sense of microarray gene expression data is the construction of genetic regulatory networks. There has been a lot of interest in looking at the interaction among genes in a holistic manner because the activity of genes are not independent of each other. Hence, network perspectives are integral to our understanding of biological interactions and to channel this insight in order to develop successful intervention methodologies. There have been a number of attempts at modeling genetic regulatory networks, such as Boolean network models [9], Differential equation models [10], Probabilistic Boolean network models [11] and Bayesian network models ([12], [13]).

One of the most important sources of biological knowledge available from years of experimental observations by biologists is in the form of signaling pathways. Signaling pathways illustrate the interactions among the various biological elements present in them. Whereas the signaling pathways available from the literature are not an accurate description of the underlying biological phenomena, they are constructed from biological experiments aimed at uncovering marginal (pair-wise) cause-effect relationships between genes involved in a biological process.

There has been a lot of recent work in integrating biological pathway knowledge with genome-level data in order to perform inferences on models that are closer representations of the actual biological system. WGCNA [14] is a package in R language for weighted correlation network analysis. WGCNA can be used to construct correlation networks using genomic data and find clusters of correlated genes. It also provides methods for defining other biologically relevant measures by using biological knowledge about genes in addition to experimental data. The algorithm SPIA

[15], is a pathway impact analysis technique that uses traditional pathway enrichment in combination with a pathway perturbation measure to facilitate pathway ranking in multiple disease datasets. PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models [16]) uses a factor graph approach to deduce pathway activities by integrating multiple types of biological data such as copy number, mRNA expression, etc. for the same disease condition. In both SPIA and PARADIGM, the focus is on detecting pathways that are more significant in a disease so that targeted therapy can be developed. Zodiac [17] is a recently proposed technique to integrate biological knowledge about genetic interactions in cancer along with experimental data to obtain an enhanced interaction map.

In this chapter, we describe a methodology that utilizes current biological knowledge from the literature to build genetic regulatory network models and integrates this knowledge with experimental genomic data using a Bayesian Network (BN) based approach. BNs are a class of directed acyclic graphs that encode independencies in a given network. BNs are a natural fit to the problem at hand since they can be used to represent causal relationships, similar to the nature of relations in biological signaling pathways. The state of each node in a Bayesian network is described by a probability distribution. A node with an outgoing edge pointing to another node is said to be a 'parent' of the latter node. Nodes that have no parents (no incoming edges) have marginal probability distributions whereas the other nodes have conditional probability distributions describing their state, conditioned on the states of their parents. Using the idea of decision making under uncertainty, we also provide a novel framework to select influential genes that govern the behavior of the network in a given environment. More specifically, our interest is in choosing genes which can be intervened with in order to manipulate pathway dynamics. This is the major difference between our method and other pathway analysis methods

mentioned earlier. While most gene selection methods are basically feature selection methods for classification or clustering, our technique is aimed at identifying genes for intervention. We use the proposed method to analyze the MAPK network in the plant defense response pathway.

The rest of the chapter is organized as follows. Section 2.2 is a brief review of some basic concepts related to Bayesian networks. In section 2.3, we elaborate on the framework and the approach to integrate experimental data into the Bayesian network framework, and also discuss the algorithm for selection of potential targets for intervention. The biological knowledge on plant-pathogen interaction is presented in section 2.4. Section 2.5 is a description of simulations and results obtained from applying our approach to plant genomic datasets. Concluding remarks are presented in section 2.6.

2.2 Modeling Bayesian Network

Bayesian network models are promising tools for the analysis of genetic regulatory networks primarily because the interactions among the components of a gene regulatory network are sparse: i.e., each gene is controlled by only a limited number of other genes, which is a very small number compared to the total number of genes in an organism. Moreover, biological systems are inherently stochastic, and the probabilistic nature of Bayesian networks gives them the ability to cope with the uncertainties involved in gene networks. A BN is a compact representation of complex relationships among a large number of random variables. This representation consists of (i) a Directed Acyclic Graph (DAG) and (ii) a conditional probability distribution for each variable, given its parents in the graph. The graph represents conditional independence properties according to which the joint probability distribution of the BN gets factorized [18].

We review some basic definitions and notations of BNs in this section. A Bayesian network is defined by a pair $\langle G, \Theta \rangle$, where G is the DAG whose nodes $X_1, X_2, ..., X_n$ represent random variables, and whose edges represent the direct dependencies between these variables. Each X_i can take values x from the domain $\mathcal{X}$. The graph G encodes Markov relationships, according to which, each variable X_i is independent of all other nodes given its parents in G. We assume that the conditional probability distributions are parameterized by the second component Θ.

The joint distribution can be decomposed into:

$$P(X_1, X_2, ..., X_n) = \prod P(X_i | Pa(X_i)) \tag{2.1}$$

where *Pa(X_i)* is the set of parents of X_i. Various methods have been proposed to infer the structure of the gene regulatory network using Bayesian network based methods ([19], [20], [21]). Some of these methods use machine learning algorithms to develop a scoring system based on features extracted from biological experimental data. But the ultimate validation of such models is carried out by comparing them with biologically relevant connections documented in a signaling pathway database. We utilize the domain expert knowledge available in the literature in the form of signaling pathways constructed by biologists to build the graph skeleton according to which the joint probability distribution of our model gets factorized.

Once we obtain the graph structure G of the Bayesian network, we proceed to the next stage of our algorithm, which is to update the model parameters using gene expression data. Though many approaches to gene expression analysis use real valued data, we use a binary framework in order to describe the state of a gene at a given time. The reason for using binary quantized gene expression values is manifold. Although the expression value of a gene is continuously varying, only certain large

variations in expression of one gene regulating the other leads to a change in expression of the gene being regulated. It is these changes that we are concerned with in this work, since our primary aim is to understand and quantify the effect of a gene on another. Working in the binary domain also offers several advantages such as noise robustness and computational simplicity. We adopt the procedure introduced by Otsu [22] to select a threshold for discretizing gene expression data. Expression values above the threshold are assigned a 1 and those below are assigned a 0.

We now proceed to illustrate the process of integration of gene expression data with the Bayesian network model. Consider a Bayesian network with N nodes. Let θ_X be the probability of success of node X , i.e., the probability that it takes value 1. Then, the probability that the node takes value 0 is $(1 - \theta_X)$. We adopt a Bayesian approach to parameter estimation, which requires us to define priors over the parameters θ_X. The prior is a probability distribution that expresses uncertainty about parameter θ_X before the data are taken into account. For each node, we choose a prior such that θ_X is beta distributed with shape parameters α_X and β_X.

$$\theta_X \sim Beta(\alpha_X, \beta_X) \tag{2.2}$$

$$Beta(\theta_X; \alpha, \beta) = C.\theta_X^{\alpha_X - 1}(1 - \theta_X)^{\beta_X - 1} \tag{2.3}$$

where C is a normalization constant, given by the reciprocal of the beta function with parameters α and β.

The prior can be assumed to have a uniform (flat) distribution before experimental data are observed. This is the case where α_X and β_X both take the value 1.

The observed data points form the likelihood in the Bayesian setting. Given the

distribution of θ_X, the observed values for a node are independent of each other. Assuming we have n observations for each node, the data likelihood follows a Binomial distribution given by:

$$P(X|Pa(X), \theta_X) \sim B(n, \theta_X) \tag{2.4}$$

$$B(k; n, \theta_X) = \binom{n}{k} \theta_X^k (1 - \theta_X)^{n-k} \tag{2.5}$$

Here, *Pa(X)* denotes the set of parents of a node X, B represents the Binomial distribution, k is the number of successes observed, $\binom{n}{k}$ is the binomial coefficient, and $\sim$ denotes that the probability on the left follows the distribution specified on the right.

We choose a beta distributed prior because the beta distribution is known to be conjugate to the binomial likelihood. Whenever we have a conjugate prior, the posterior distribution belongs to the same family of distributions as the prior [23]. Hence, the conditional posterior probability distributions of the nodes are beta distributed with shape parameters α'_X and β'_X.

$$P(\theta_X|X) \sim Beta(\alpha'_X, \beta'_X) \tag{2.6}$$

where $\alpha'_X = (\alpha_X + k)$; $\beta'_X = (\beta_X + n - k)$; k is again the number of one's observed among data points of node X. For a Beta distribution, the expected value is given by,

$$E(\theta_X|X) = \frac{\alpha'_X}{\alpha'_X + \beta'_X} \tag{2.7}$$

For the purpose of illustration, let us consider the Bayesian network shown in Fig.2.1, with two binary-valued nodes A and B. Let the probability of success parameter θ_A of node A, and $\theta_{B=1|A=1}$ and $\theta_{B=1|A=0}$ of node B, have prior probability distributions that are Beta distributed with shape parameters (1,1). Assume that

the observed data for the two nodes are as shown in Table 2.1.

FIG. 2.1. A simple Bayesian network with two nodes A and B, connected such that A is the parent of B. Figure reprinted with premission from [137].

Then, the posterior probability distribution of θ_A is given by:

$$P(\theta_A) \sim Beta(\alpha'_A, \beta'_A) \tag{2.8}$$

where $\alpha'_A = n_{11} + n_{10} + 1$ and $\beta'_A = n_{01} + n_{00} + 1$. Here, n_{11} is the number of observations such that $A = 1$ while $B = 1$, n_{01} is the number of observations such that $A = 0$ while $B = 1$, etc.

In a similar manner, the conditional posterior probability distribution of $\theta_{B=1|A=1}$ is given by:

$$P(\theta_{B=1|A=1}) \sim Beta(\alpha'_{B_1|A_1}, \beta'_{B_1|A_1}) \tag{2.9}$$

where $\alpha'_{B_1|A_1} = n_{11} + 1$ and $\beta'_{B_1|A_1} = n_{10} + 1$.

As more data are observed, we can update the values of α's and β's so that the posterior probabilities approach the true underlying distribution. Note that it is

TABLE 2.1. Sample observations for binary valued random variables A and B. Table reprinted with permission from [137].

Node A	0	1	0	1	0	1	1	0	1
Node B	1	1	1	1	0	1	1	0	0

possible to use other priors for the parameters, and in such cases, computationally efficient estimation techniques (Markov Chain Monte Carlo (MCMC) methods, for instance) may need to be used for estimating posterior probability distributions.

2.3 Identifying Significant Genes

Given a genetic regulatory network, it is important to differentiate between those genes that have a major influence on the regulation of the child gene and those that only have a minor influence. Biologically, a gene with stronger influence can overshadow the effect of other genes that have minimal influences. Many such biological relationships are known to exist. For instance, the activation of the p53 gene, which is a well-known tumor suppressor, actively leads to the expression of various genes that promote apoptosis, whereas p73, another tumor suppressor belonging to the same group of signaling pathway elements is less effective in activating apoptotic genes [24].

Our objective is to maintain some downstream reporter nodes in the network at certain desirable states. In such a scenario, we have different options for the choice of point(s) of intervention to prod the network towards the specific behavior which we are interested in. For each gene, we have an associated probability distribution over its possible states and its influence on a desired node in the network. Therefore, the gene selection problem is essentially a problem of decision making under uncertainty. In order to find optimal decisions, we assign numerical utilities to all possible outcomes and then choose the decisions that result in maximal utility value.

Utility is a subjective notion and the design of the utility function depends on the objective of the action (gene intervention in our case) and the nature and preference for tools available to cause the action. For instance, in the case of gene intervention, gene knockouts may be easier than gene activation, and have stronger downstream effects and hence have more utility to the biologist, depending on the type of genetic network under consideration. We illustrate the decision making process using an example.

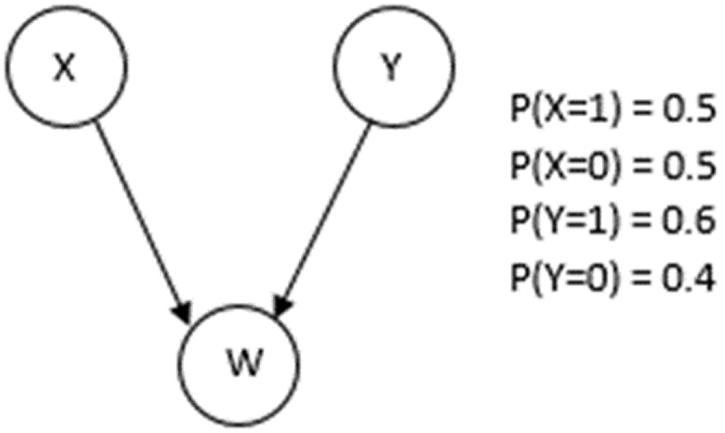
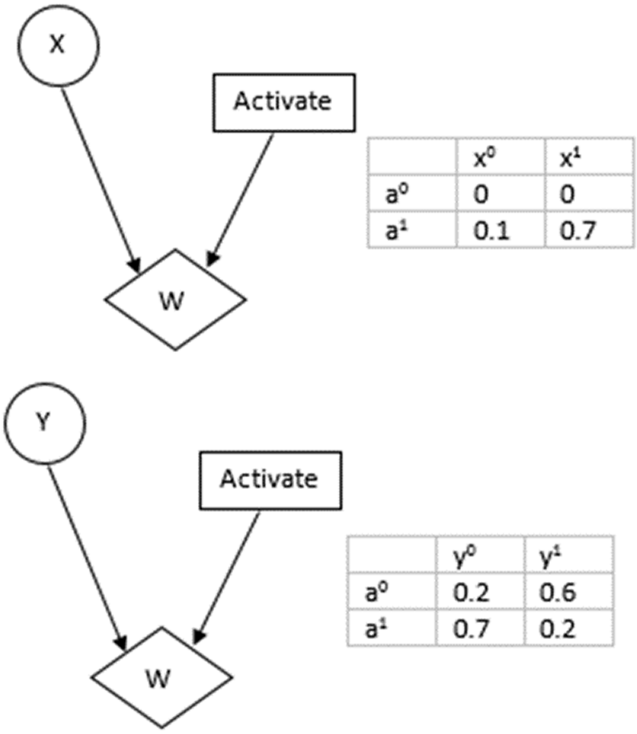


FIG. 2.2. Example Bayesian Network. The marginal probabilities of binary valued nodes X and Y are shown in the table alongside. Figure reprinted with permission from [137].

Example: Consider the Bayesian network shown in Fig. 2.2. We assume that all nodes have binary states. We draw an *influence diagram* from the given Bayesian network, as shown in Fig. 2.3, to represent the decision making scenario. In addition to the original random variables in the network (the choice variable), we have the decision variable (rectangular node) and utility variable (diamond shaped node). We assume that the decision variable 'Activate' can take binary values, 1 or 0, corresponding to activation (force the choice node to be in state 1) or inhibition (state 0). Therefore, we have four different options for achieving the goal: Activate X, Inhibit X, Activate Y, or Inhibit Y. The utility variable (W in this example) represents the utility obtained from making the decision, and is a deterministic function of X (or Y) and the decision variable Activate. Thus, for each combination of the parent nodes of W, this function specifies a real-valued utility. The utility functions of nodes X and Y are shown as tables in Fig. 2.3. Here, a^1 represents the decision to activate the gene and a^0 represents the decision to inhibit the gene.



	x^0	x^1
a^0	0	0
a^1	0.1	0.7

	y^0	y^1
a^0	0.2	0.6
a^1	0.7	0.2

FIG. 2.3. Influence diagrams and tables showing utility values of node X and node Y corresponding to different states of node 'Activate': a^1 and a^0. Variables X and Y each have two possible states x^0, x^1, and y^0, y^1, respectively. Figure reprinted with permission from [137].

Fig. 2.4 is the corresponding *decision tree* for the scenario in the example. A decision tree is nothing but a graphical representation of all possible scenarios that might be encountered in a decision problem, and the corresponding outcomes and their utility values. A common approach to arrive at profitable decisions in the decision tree is to employ the *backward induction* algorithm. The algorithm moves backward in the tree, beginning from the leaves, computing the maximum expected utility (MEU) achievable at each node. The MEU at a leaf is the utility value associated with that leaf's outcome. As we move up the tree, if we encounter a nature node (i.e., not a decision node), then the MEU is the weighted average of the

expected utilities at each of the node's children, where each child's utility is weighted according to the distribution defined by nature over the node's children. If the node encountered is a decision node, then a decision is made to select the child whose MEU is largest. The expected utility for each action is given by:

$$EU[X, A = a^1] = \sum_x P(A = a^1 | X = x) p(X = x) \tag{2.10}$$

$$EU[X, A = a^0] = \sum_x P(A = a^0 | X = x) p(X = x) \tag{2.11}$$

In this example, the maximum expected utility for node X is given by:

$$\begin{aligned} MEU[X, A = a] = \max_a [&((0.7)(0.5) + (0.1)(0.5)), \\ &((0)(0.5) + (0)(0.5))] = 0.40 \end{aligned} \tag{2.12}$$

and the maximizing action is a^1. Whereas, node Y has a maximum expected utility of:

$$\begin{aligned} MEU[Y, A = a] = \max_a [&((0.2)(0.6) + (0.7)(0.4)), \\ &((0.6)(0.6) + (0.2)(0.4))] = 0.44 \end{aligned} \tag{2.13}$$

and the maximizing action is a^0. Hence inhibition of Y is the preferred decision in order to maximize the expected utility.

The principle of maximum expected utility and the backward induction approach have been discussed amply in the literature ([26], [25]). We proceed to use these techniques in the plant MAPK network where the goal is to find important lever genes that can be used to manipulate the global behavior of the defense response network. We consider two important factors in the design of the utility function: the first factor is the effect of the gene on the utility variable; and second, the fact that

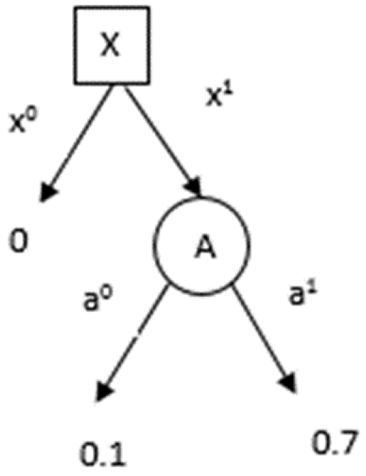


FIG. 2.4. Decision tree constructed from the influence diagram, illustrating the process of selecting the most advantageous action for node X, using the backward induction algorithm. Figure reprinted with permission from [137].

the effort involved in flipping a gene depends on its predisposition to be activated or inactivated, given that other genes that influence it are in a certain state. A naïve way to determine the best set of genes is to exhaustively score each possible combination of genes and pick the set with the maximum utility. However, one can devise a better algorithm by making use of the recursive nature of the computations for determination of the utility value. The algorithm provided is a stepwise description of our methodology that makes use of maximum expected utility to select points of intervention in the genetic regulatory network.

Algorithm Algorithm for selecting significant genes

Input: Bayesian network graph G, Nodes $X_1, X_2, ..., X_n$, conditional probabilities, desired state of reporters
Output: Nodes, Control actions
State of nodes: 0 or 1
Define Utility function:

$$U(X, a) = u1(X, a)u2(X) \tag{2.14}$$

▷ $u1(X, a)$ is the utility obtained from taking an action 'a' that enables gene X to induce the reporter node to be at the desired state.
▷ $u2(X)$ is the utility associated with gene X being in a certain state, given that it's ancestors are in some predefined states.
Step 1: Calculate $u2(X)$: Start at the top of graph G. For root nodes, $u2$ is the marginal probability distribution. Store $u2$. Proceed one level down in the graph, marginalize out parent node from child and parent's joint probability distribution to get $u2$ for the child node. Number of computations is reduced by using a dynamic programming approach: $u2$ of parent node calculated in each iteration is used in calculating child's marginal in next iteration. Continue till all nodes are exhausted.

Step 2: Calculate $u1(X, a)$: Utility of each node in inducing desired state of reporter node. This is given by the conditional probability distribution of the reporter node conditioned on the states of each of the other nodes. Store $u1$ for each node.

Step 3: Calculate $U(X, a)$ for each node and action. Choose those nodes that maximize the expected utility.

2.4 MAPK Cascade in Plant Defense

Plants are immobile organisms that rely on their innate immunity to prevent harmful microorganisms such as fungi, oomycetes, bacteria and viruses from invading them and causing infectious diseases. Plant surface structures such as spines, thorns, and trichomes form the first line of defense against potentially destructive organisms. In addition to these constitutive defenses, plants also possess inducible defenses that are triggered in response to attack. The plant immune system is composed of sensors that can recognize microbial molecules such as chitin, lipopolysaccharides (LPS), peptidoglycan, flagellin, elongation factor Tu (EF-Tu), etc., collectively termed as Microbe-Associated Molecular Patterns (MAMPs) and respond accordingly by altering gene expression and metabolism to localize the invasion of the pathogen. These sensors, or pattern recognition receptors (PRRs), are essentially receptor-like kinases (RLKs) and receptor-like proteins (RLPs) that are in a rest state prior to ligand binding. When certain microbial molecules bind to the PRRs, a number of downstream plant defense response genes are activated, and the phenomenon is known as MAMP-triggered immunity [27].

Well-adapted microbial pathogens have developed mechanisms to elude detection by RLKs, thereby evading the induction of primary defense responses. The secretion system of bacteria plays a major role in this process. Molecules called effectors, injected by the bacterial secretion system into the host plant cells are capable of suppressing the primary defense response of plants [29]. Plants in turn have evolved a secondary defense mechanism to defend themselves against such potent pathogens, based on a class of specialized Resistance genes (R genes) that monitor the pathogen-injected intracellular effectors and activate a cascade signaling that leads to defense response.

The eventual transcription of major plant defense genes is regulated by the Mitogen Activated Protein Kinase (MAPK) pathway. MAPKs are essentially phosphorylating enzymes. They are organized in protein complexes or modules. MAPK pathways are highly conserved, and they are prime regulators of proliferation and stress response in all eukaryotes. A MAPK cascade consists of a MAPKKK-MAPKK-MAPK module. The compartmentalization of MAPK modules and targets of activated MAPKs brings specificity to MAPK signaling [31].

In this paper, we focus on the MAPK network and the genes involved in the signaling cascade of the MAPK pathway specific to plant defense response. Though high throughput arrays provide measurements for thousands of genes, interactions among genes in small important networks are believed to have major impacts on disease conditions, which is another reason why we focus on the MAPK subnetwork of the plant defense response. We use a Bayesian network model of the MAPK pathway and perform inference on this model as described in section 2 to update the parameters of this model. The KEGG pathway database ([33], [32]) is a repository of molecular interactions and serves as a gold standard for biological signaling pathways. The structure of the MAP Kinase cascade, shown in Fig. 2.5, is obtained from the set of plant-pathogen interactions reported in the KEGG database for the model plant species *Arabidopsis Thaliana*.

2.5 Simulations

In this section, we discuss the application of our method to select important points of intervention in the plant defense response to pathogens. We specifically look at the interaction of the plant *Arabidopsis Thaliana* with bacterial pathogens. Arabidopsis is widely used as a model plant since it is easily manipulated, genetically tractable, and a lot of knowledge is available about the behavior of the plant under various stress conditions. We use experimental data obtained over multiple experiments and deposited in the publicly accessible NCBI database [34]. The results obtained from the application of our approach to these datasets are interpreted in a biologically meaningful way.

2.5.1 Data Sets

We use real data sets to test the performance of our Bayesian network model. Three gene expression series data sets GSE17464, GSE19109, and GSE18978 were obtained from the NCBI GEO database. Each of these datasets contain gene expression data obtained from experiments where the Arabidopsis plant was exposed to bacterial molecules and the gene expression changes induced by this stimulus were measured using microarrays.

2.5.2 Bayesian Network Estimation

The Bayesian network graph of the plant-pathogen interaction pathway is as shown in Fig. 2.5. The parameters that encode the probability distributions of nodes in the network are shown in Fig. 2.6. The expression values of the nodes were extracted from the datasets using R Bioconductor software [35] and discretized. The datasets were pooled to generate the sequence of observations used to update the parameters of the Bayesian Network.

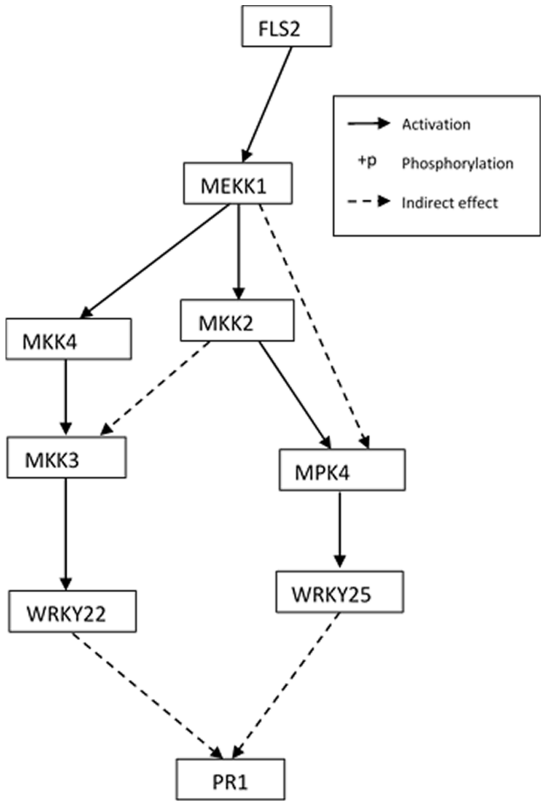


FIG. 2.5. Structure of the Mitogen Activated Protein Kinase cascade. The cascade culminates in the activation of plant defense response gene Pathogenesis-related protein 1 (PR1). Figure reprinted with permission from [137].

We ran simulations for analyzing the behavior of our model. Initially, we take the prior for all nodes to be $Beta(1, 1)$, which is a uniform distribution over the finite interval $[0, 1]$. Such a prior provides equal weights to all possibilities in the parameter space, and is hence noninformative. Using (6), we proceed to obtain the posterior distribution by updating the prior with the gene expression data.

The expectations of the updated posterior probabilities are tabulated in Table 3.2. Here, the means shown are the expected values of each state of each node conditioned on each possible combination of states of its parent nodes. For instance, the expec-

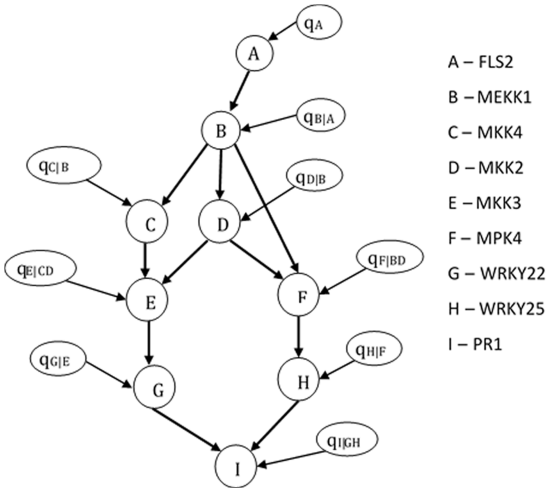


FIG. 2.6. Bayesian Network model illustrating the conditional dependencies in our network. The parameters q_A, $q_{B|A}$ etc. of the network are estimated using a dynamic programming approach as described in the algorithm. Figure reprinted with permission from [137].

tation value $\overline{B_1}|A_1$ is calculated using (7) :

$$E(\theta_B = 1|A = 1) = \frac{\alpha'_{B_1|A_1}}{\alpha'_{B_1|A_1} + \beta'_{B_1|A_1}} \tag{2.15}$$

2.5.3 Selection of Points of Intervention

The Bayesian network and the conditional probability table associated with the Bayesian network are used to select significant nodes in the network. Let us consider that our goal is to achieve transcription of the defense response gene Pathogenesis-related protein 1 (PR1). This is the leaf node of the network shown in Fig. 2.5. Since we want PR1 (node I) to be in an activated state, we associate a utility value of 0 with PR1 being in state 0. Then, for each node in the network, we use the approach explained in section 3 to compute the expected utility value and select the nodes with maximal expected utility to be the points of intervention. In this work,

TABLE 2.2. The means of beta posterior conditional probability distributions. Table reprinted with permission from [137].

Node	Mean
$\overline{A_1}$	0.82
$\overline{B_1}\|A_1$	0.9
$\overline{B_1}\|A_0$	0.33
$\overline{C_1}\|B_1$	0.901
$\overline{C_1}\|B_0$	0.66
$\overline{D_1}\|B_1$	0.9
$\overline{D_1}\|B_0$	0.3
$\overline{E_1}\|D_1C_1$	0.82
$\overline{E_1}\|D_1C_0$	0.5
$\overline{E_1}\|D_0C_1$	0.66
$\overline{E_1}\|D_0C_0$	0.3
$\overline{F_1}\|D_1B_1$	0.9
$\overline{F_1}\|D_1B_0$	0.5
$\overline{F_1}\|D_0B_1$	0.5
$\overline{F_1}\|D_0B_0$	0.3
$\overline{G_1}\|E_1$	0.82
$\overline{G_1}\|E_0$	0.5
$\overline{H_1}\|F_1$	0.9
$\overline{H_1}\|F_0$	0.66
$\overline{I_1}\|G_1H_1$	0.9
$\overline{I_1}\|G_1H_0$	0.5
$\overline{I_1}\|G_0H_1$	0.66
$\overline{I_1}\|G_0H_0$	0.5

we only consider single interventions. It is also possible to think of scenarios with combinations of interventions. Fig. 2.7 is a graphical representation of the utility obtained through distinct interventions at each of the nodes in the network with the goal of maintaining the reporter node PR1 at 1. The first bar in the graph shows the utility associated with activation of gene FLS2, and the second bar is the utility obtained on inhibition of gene FLS2. The rest of the bars show utilities for each of the other genes, on activation and inhibition. In this network, activation of the network elements has more utility value than their inhibition, as seen from the graph.

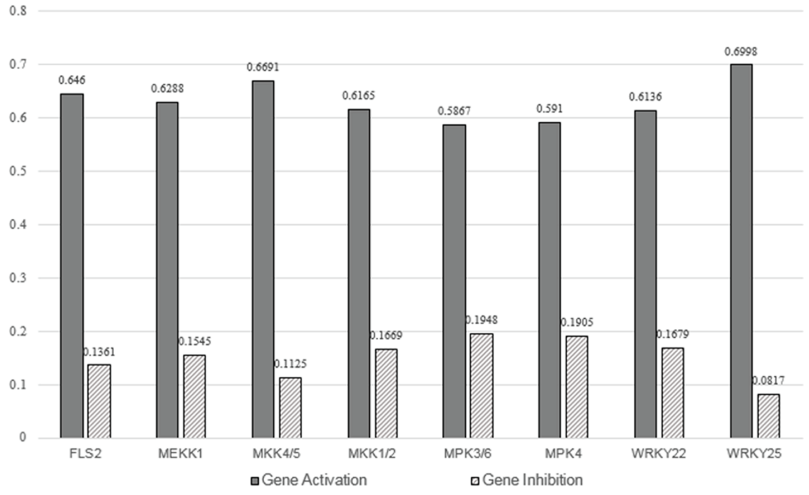


FIG. 2.7. Bar graph illustrating the utility obtained by intervening with different nodes in the network to achieve desired state of node representing the gene Pathogenesis-related protein 1 (PR1). Figure reprinted with permission from [137].

2.5.4 Results and Discussion

The inference from the application of the gene selection algorithm is that WRKY25 is the most preferred node for intervention. WRKY25 belongs to the WRKY I group of proteins. These proteins are involved in regulation of certain processes including

pathogen defense, and response to various stresses. WRKY25 is an important regulator of biotic stress and is induced by salicylic acid. Recent studies have shown that in Arabidosis thaliana WRKY25 and WRKY39 are positive regulators in thermotolerance [36]. Hence, WRKY25 is a potent point of intervention in the plant pathogen response system. The second most significant point of intervention according to our algorithm is MKK4. Indeed, this important member of the MAPK cascade is vital in plant immunity, and over-expression of MKK4/5 has been seen to increase resistance to powdery mildew in wild type Arabidopsis [37].

2.6 Concluding Remarks

MAPK cascade is an important point of convergence of abiotic and biotic stress stimuli in plants. In this chapter, we developed a Bayesian network based approach for modeling and inference of the MAP Kinase cascade in plants, which uses the graph structure obtained from signaling pathways in the biology literature.

We demonstrated the integration of biological knowledge from the literature with real gene expression data from the NCBI repository. Subsequently, we have made an attempt to quantify the effect of a gene on another and used this influence as a tool to drive the network behavior towards a desired output. Although the demonstration of the working of our algorithm was carried out using gene expression data, the method can work with other types of genomic or proteomic data or a combination of multiple types of data. An important consideration in our approach is that biologists are always looking for one crucial lever gene that can be manipulated in real experiments rather than a bunch of genes that have some influence on the desired response, since it is both time-consuming and expensive to test the effect of manipulation of multiple genes on the gene network.

The approach presented in this chapter for selection of points of intervention

can be applied to any Bayesian network that represents biological signaling in an organism. The method is applicable to networks of any size provided there are no cycles in the network. From the simulation results discussed above we see that intervention with genes that have greater downstream effects are more likely to result in global network changes. A variety of factors can be considered to score the gene selection. The ease with which a gene can be maintained in a certain state given the state of its parent genes is the most important factor, which we have considered in the algorithm. Another possible factor could be the relative proximity between the target and the manipulated gene in the genetic regulatory network. The central idea here is that some 'important' genes have the potential to lead the network to a certain desired state. This desired objective can be attained with lesser effort using a properly chosen gene as compared to another gene which may not have as strong an influence.

The gene selection algorithm proposed in this paper is sensitive to the structure of the network. Incorrect connections in the network structure may lead to the selection of genes that are not very influential in driving other genes. Combining multiple sources of biological data may help in improving the robustness of the method.

We used a beta-binomial model for representing the interactions among nodes in the MAPK pathway. However, as we mentioned earlier, it is possible to use other parametric models which make use of nonconjugate priors in the modeling. For such distributions, MCMC methods can be utilized to sample from the posterior distributions and obtain kernel density estimates. In cases where the form of the posterior is unknown, algorithms such as the Metropolis Hastings algorithm can be employed in addition to MCMC to obtain estimates of the posterior distribution. This study was focused on learning one graph that corresponds to a specific condition for the genes. However, it is also possible to analyze an ensemble of graphs that correspond to

different diseased conditions, and compare such models to investigate critical nodes that can be targeted for improving disease resistance. Further study is required to analyze the development of such models. Integration of other types of genomic, proteomic and sequencing data into the Bayesian network will also be explored in future work. The formulation of utility functions that are custom-designed according to the biological intervention being considered is another direction for future studies.

3. TRANSCRIPTIONAL MODIFICATIONS IN *Solanum lycopersicum* DUE TO COMBINED EFFECT OF DROUGHT AND PATHOGEN STRESS

3.1 Introduction

RNA-Seq technology has revolutionized the field of transcriptome analysis by providing precise measurements of transcript levels in a high throughput manner. RNA-Seq uses deep-sequencing technologies to list out all the different species of transcripts and to quantify the changing expression levels under different experimental conditions (Tarazona, Garcia-Alcalde, Dopazo, Ferrer, & Conesa, 2011). When compared to DNA microarrays, RNA-Seq has low background signal, high resolution and higher sensitivity (Wang, Gerstein, & Snyder, 2009). RNA-Seq has the added advantage of unambiguously mapping DNA sequences to specific regions of the genome. It also provides a large range of expression levels over which we can detect transcripts. (Wang, et al., 2009). In this chapter, we discuss results obtained from computational analysis of RNA sequences of the transcriptome of Solanum lycopersicum subjected to multiple stress conditions. The sequences were obtained from Illumina sequencing (Bennett, 2004) of RNA extracted from tomato plant samples.

Being sessile organisms, plants cannot use locomotion to defend themselves from unfavorable conditions. Plants have developed innate defense mechanisms to detect and respond to biotic stresses such as attack by herbivores and pathogens (Boari & Malone, 1993; Bowles, 1990) and abiotic stresses such as drought (Holmstrom, Mantyla, Wellin, & Mandal, 1996) and high salinity (Yamaguchi-Shinozaki & Shinozaki, 1994). In fact, the plant signaling molecules that play key roles in plant response to different kinds of stresses are characteristically different (Chinnusamy, Schumaker, & Zhu, 2004; Glazebrook, 2001; Reymond, Weber, Damond, & Farmer,

2000; Thomma, Penninckx, Cammue, & Broekaert, 2001). There have been multiple studies aimed at investigating the characteristics of plant defense responses to combined stress conditions that indicate plants exhibit specific responses that are different at the molecular level while combating simultaneous stress conditions (Atkinson & Urwin, 2012; Kunkel & Brooks, 2002; Suzuki, Rivero, Shulaev, Blumwald, & Mittler, 2014). More interestingly, evidences suggest that plant responses to isolated stress conditions cannot be used in inferring the responses to combined stress conditions (Jones, Flowers, & Jones, 1989; Pandey, Ramegowda, & Senthil-Kumar, 2015; Rejeb, Pastor, & Mauch-Mani, 2014). In other words, the principle of superposition, typical of linear relationships, does not hold.

Drought is an important stress factor that affects the yield of plants by interfering with their normal functioning (Farooq, Wahid, Kobayashi, Fujita, & Basra, 2009; Morgan, 1984). There have been numerous studies (Easlon & Richards, 2009; Giunta, Motzo, & Deidda, 1993; Taji et al., 2002) into the effects of drought in various plant species, including tomato, which is a major food crop in many parts of the world. Studies have revealed multiple gene transcriptions in tomato induced by water stress that help the plant survive this unfavorable condition (Gong et al., 2010; Orellana et al., 2010; X. Zhang et al., 2011).

Bactericera cockerelli (Sulc), a psyllid commonly found on potato and tomato crops, is a serious pest that feeds on the phloem of tomato plants. It is also the vector of the bacterium 'Candidatus Liberibacter Solanacearum' (Crosslin, Lin, & Munyaneza, 2011). Candidatus Liberibacter Solanacearum (Lso) is a phloem-restricted gram negative bacterium that causes the Zebra Chip disease in potato (Brown, Rehman, Rogan, Martin, & Idris, 2010; Garon-Tiznado et al., 2009; Munyaneza, 2012). A number of studies have been carried out to investigate the response of tomato plants to psyllid herbivores (Liu, Johnson, & Trumble, 2006) and bacterial

pathogens (Casteel, Hansen, Walling, & Paine, 2012; Li & Steffens, 2002; Pedley & Martin, 2004). Transcriptomics can also be used to understand tri-trophic interactions. The study of plant responses to psyllids and to psyllids and Lso revealed that plants respond to these two challenges differently. In particular, when Lso is present, plant response might be delayed (Ordom et al, in revision).

In nature, plants are confronted to a multitude of biotic and abiotic attackers. Our interest is to study the response of tomato plants to a combinatorial stress condition involving herbivory by Bactericera cockerelli, infection by Lso, and water shortage.

3.2 Materials and Methods

3.2.1 Plant and Insect Material

Solanum lycopersicum cv. Money Maker (Thompson & Morgan) seeds were planted in Sun Gro®Metro-Mix 900 soil. Plants were grown under L16:D8 (light:dark) cycles and were provided adequate water and fertilizer (Miracle-Gro ®Water Soluble Tomato Plant Food, 24-8-16 NPK) following recommendations. Laboratory B. cockerelli not harboring Lso (not infected) and harboring Lso (Lso-infected) were maintained on S. lycopersicum cv. Money Maker in 14" X 14" X 24" insect cages (BioQuip) at room temperature and a L16:D8 photoperiod. Diagnostic PCRs were routinely performed regularly to test for Lso infection (Nachappa et al. 2012). No psyllid from the uninfected colony tested positive for the presence of Lso, while on average, over 90% of insects from the Lso-infected colony tested positive for Lso.

3.2.2 Treatments

Four-week-old plants were individually transplanted into 3.5-inch square pots with dry soil and were subjected to one of two water treatments: (1) 200 mL water weekly (control), or (2) 50 mL water weekly (water-stressed) (Huot and Tam-

TABLE 3.1. Experimental Setup

Sample	Water Regime	Treatment
SCD-1	50ml	Control
SCD-2	50ml	Control
SCD-3	50ml	Control
SPD-1	50ml	Psyllid herbivory
SPD-2	50ml	Psyllid herbivory
SPD-3	50ml	Psyllid herbivory
SLD-1	50ml	Lso post Psyllid herbivory
SLD-2	50ml	Lso post Psyllid herbivory
SLD-3	50ml	Lso post Psyllid herbivory
SCC-1	200ml	Control
SCC-2	200ml	Control
SCC-3	200ml	Control
SPC-1	200ml	Psyllid herbivory
SPC-2	200ml	Psyllid herbivory
SPC-3	200ml	Psyllid herbivory
SLC-1	200ml	Lso post Psyllid herbivory
SLC-2	200ml	Lso post Psyllid herbivory
SLC-3	200ml	Lso post Psyllid herbivory

borindeguy, in revision). All experiments were conducted at room temperature ($\sim$ 23 $^\circ$ C) under L16:D8 photoperiod. A week after the instauration of the water treatment, plants were assigned to one of three psyllid treatments: (1) no insects, (2) 10 3rd instar nymphs from the uninfected colony, or (3) 10 3^{rd} instar nymphs from the Lso-infected colony. Insects were placed in organza pouches. Control plants were mock-infested. One week after the infestation and two weeks after water treatment instauration, the top-most fully expanded leaf was collected, flash-frozen and kept at -80 C. In total there were six treatments (two water regime treatments and 3 insect treatments), and three biological replicates per treatment (Table 3.1).

After leaflet collection, the leaf with the psyllids was removed and plants were kept for three additional weeks and tested for Lso infection by diagnostic PCR as in (J. Levy, A. Ravindran, D. Gross, C. Tamborindeguy, & E. Pierson, 2011). Forward and

reverse primers targeting Lso 16S ribosomal RNA gene were used for Lso detection (P. Nachappa, Levy, Pierson, & Tamborindeguy, 2011) and primers targeting tomato elongation factor 1 (EF1) (Punya Nachappa, Levy, Pierson, & Tamborindeguy, 2014) were used as internal control.

3.2.3 RNA Extraction

RNA was purified using Trizol (Thermo Fisher Scientific) following the manufacturer's instructions. DNA contamination was removed using TURBO DNA-freeTM Kit (Life Technologies, Carlsbad, CA). The 18 samples were submitted to the AgriLife Genomic and Bioinformatic Center. Libraries were made following the TruSeq RNA sample preparation protocol. Sequencing of libraries was performed using three lanes of the Illumina SE 110 bp using the HiSeq 2500 platform. We use computational tools to analyze genes that are differentially expressed under different stress conditions, and signaling pathways that are significantly enriched, in order to understand transcriptional modifications resulting from this mixture of stress conditions in tomato.

3.2.4 Gene Expression Quantification

RNA-Seq reads were obtained through Illumina sequencing of 18 different library samples. The sample descriptions are as provided in Table 3.1.

RNAseq reads were mapped to the *Solanum lycospersicum* genome version 2.5 (Solgenomics) using Tophat2 (Trapnell et al., 2009) with standard parameters. The bam files obtained from this mapping were used to count the number of reads belonging to each gene of Solanum lycopersicum using the SummarizeOverlaps feature of the Genomic Alignments software package (Lawrence et al., 2013) in R (Team, 2014). Finally, the DESeq2 software package, was used to compute differential expression. (Love, Huber, & Anders, 2014).

3.2.5 Data Analysis

Differential gene expression analysis was carried out using the DESeq2 Bioconductor package in R (Anders & Huber, 2010; Love, et al., 2014). We performed the following comparisons among treatments: Control vs Drought, Control vs Bactericera cockerelli, Bactericera cockerelli vs Bactericera cockerelli plus Drought, Control vs Bactericera cockerelli plus Lso, and Lso plus Bactericera cockerelli vs Lso plus Bactericera cockerelli plus Drought. DESeq2 fits a generalized linear model (GLM) to the data, where a negative binomial distribution is used to model the counts per gene and sample. Wald test (Kodde & Palm, 1986) was performed to determine the statistical significance of the differences in gene expression. Genes were taken to be differentially expressed if the p-value was found to be $<= 0.05$ (after implementing Bonferroni correction) and a log2-fold change $<= -1$ or $>= 1$.

3.2.6 GO Term Enrichment Analysis

We used the EnrichmentBrowser package in R (Geistlinger, Csaba, & Zimmer, 2016), augmented with in-house R scripts, to identify enriched gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets in the set of differentially expressed genes obtained under different experimental conditions.

3.3 Results

Through Illumina Sequencing, we obtained about 1.9 billion reads with an average read length of 96 base pairs. About 85.23% of the raw reads were successfully mapped to the tomato genome using Tophat2. Each condition in the experimental setup was represented by three biological replicates, resulting in a total of eighteen samples. Table 3.2 provides a summary of the sequenced reads for the various samples and the reads that were mapped to the tomato genome. Here, SCD-1 refers to Solanum

TABLE 3.2. Number of Reads Sequenced and Mapped with TopHat2

Sample	Sequenced reads	Uniquely mapped reads	Multiple mappings	Alignment rate
SCD-1	9461487	7950880	205194	86.2
SCD-2	8990862	7537957	220316	86.3
SCD-3	9799054	8143014	213927	85.3
SPD-1	9033976	7732586	197445	87.8
SPD-2	10951863	9353556	256857	87.8
SPD-3	11944446	10189364	294604	87.8
SLD-1	10765031	8308824	206923	79.1
SLD-2	9860230	7927115	198675	82.4
SLD-3	11128390	9121668	261289	84.3
SCC-1	8405996	6815428	173284	83.1
SCC-2	10974756	8976640	221305	83.8
SCC-3	9212089	7493711	281455	84.4
SPC-1	13431753	10980919	286688	83.9
SPC-2	14644454	12607231	342516	88.4
SPC-3	12642467	10718418	268584	86.9
SLC-1	11447923	9458807	243414	84.8
SLC-2	10822254	9051711	219217	85.7
SLC-3	8719671	7332567	186401	86.2

lycopersicum sample under Control condition (Lso-negative), under Drought (50 ml water regime), while SCD-2 refers to a replicate of the sample under the same conditions, and so on. Table 3.1 provides descriptions of all eighteen samples used in the study.

3.3.1 Differential Expression of S. lycopersicum Genes in Response to Psyllid Herbivory Preceded by Exposure to Drought

A comprehensive analysis revealed that, out of 24149 genes with nonzero read count, 379 genes were up-regulated and 322 were down-regulated. Further analysis revealed that, among the up-regulated genes, about 34% of the genes were found to be associated with GO terms in the the Molecular Function class, 41% were associated with GO terms in the Biological Process class, and 25% belong to the Cellular

Component class. When the tomato plants were sufficiently watered, psyllid herbivory conditions led to up-regulation of a number of genes such as Thaumatin-like protein, Polygalacturanose inhibitor protein and Leucine-rich repeat receptor-like protein kinase PEPR2, which are involved in plant defense (Datta et al., 1999; Roux et al., 2011). Genes down-regulated under psyllid herbivory (SPC vs SCC) include Urea active transporter-like protein, Kunitz trypsin inhibitor, associated with herbivore defense in poplar (Major and Constabel 2008), Chitinase-like protein, associated with altering root structure in response to environmental conditions (Fujimoto, Ohta, Usui, Shinshi, & Ohme-Takagi, 2000; Hermans, Porco, Verbruggen, & Bush, 2010; Major & Constabel, 2008), Ethylene-responsive transcription factor 2 (ERF2), involved in disease resistance pathways (Fujimoto, Ohta et al. 2000).

Differential expression analysis of the combined stress condition involving water shortage and psyllid herbivory (SPD vs SPC) revealed 10% more genes up-regulated as compared to the previous condition. Among the differentially expressed genes, 418 were up-regulated and 305 were down-regulated. Among the up-regulated genes, about 34% of the genes belong to the Molecular Function class, 49% belong to the Biological Process class, and 17% belong to the Cellular Component class. A summary of the up-regulated genes and their functions is provided in Table 3. Interestingly, some of the genes down-regulated in the combined stress condition are linked to plant defense to insect pathogens, such as Leucine-rich repeat receptor-like protein kinase PEPR2 (Roux, Schwessinger et al. 2011), Xylanase inhibitor, implicated in defense response in wheat (Igawa et al., 2004), Chitinase, associated with plant defense against insect pathogens (Ding et al., 1998), and NPR1-like protein, associated with Salicylic acid (SA) - mediated disease resistance (Fan & Dong, 2002).

3.3.2 Differential Expression of S. lycopersicum Genes in Response to Lso Infection and Psyllid Herbivory Preceded by Exposure to Drought

When the tomato plants were provided sufficient water, Lso bacterial infestation post psyllid herbivory (SLC vs SCC) resulted in the up-regulation of 282 genes and down-regulation of 335 genes out of a total of 24168 genes. Among the up-regulated genes, about 39% of the genes belong to the Molecular Function class, 44% belong to the Biological Process class, and 17% belong to the Cellular Component class. Up-regulated genes include BZIP Transcription family protein, associated with DNA binding (Izawa, Foster, & Chua, 1993), Pectate lyase, associated with disease resistance to pathogens (Vorwerk, Somerville, & Somerville, 2004), Protein BREVIS RADIX, involved in response to Abscisic Acid (ABA) (Rodrigues et al., 2009), and RLK, receptor like protein, a putative resistance protein with an antifungal domain (Shiu & Bleecker, 2001). Down-regulated genes include Kunitz trypsin inhibitor, Major allergen Mal d 1, associated with defense response to biotic stimulus, Cathepsin B-like cysteine proteinase, associated with intracellular protein degradation, and JAZ (Jasmonate ZIM-domain protein), associated with response to insect attack in plants (Chung et al., 2008).

In the case where the tomato plants were subjected to all the stresses (ie., water shortage and Lso infection post psyllid herbivory), out of 24028 genes with nonzero read count, a total of 683 genes were found to be significantly differentially expressed. Among these, 314 genes were up-regulated under the combined condition (SLD) and 369 were down-regulated. Among the up-regulated genes, about 24% of the genes belong to the Molecular Function class, 52% belong to the Biological Process class, and 24% belong to the Cellular Component class. Genes up-regulated under the combined stress condition were Beta-1,3-glucanase, associated with abiotic stress re-

sponse and pathogenesis-related family of proteins (Wu & Bradford, 2003) , Aspartic proteinase-2, associated with digestive elements in pitcher fluids of carnivorous plants (An, Fukusaki, & Kobayashi, 2002), Flavanone 3-hydroxylase, known to be involved in the flavonoid biosynthetic process (Shen et al., 2006), Calcium-dependent protein kinase 2, associated with plant defense response (Romeis, Ludwig, Martin, & Jones, 2001), Polygalacturonase-like protein associated with cell wall biogenesis/ degradation (Roux, Schwessinger et al. 2011), Water-stress inducible protein 3, associated with response to water stress (Shinozaki & Yamaguchi-Shinozaki, 1997).

Genes down-regulated under the drought combination include WRKY78, associated with stem elongation and seed development in Oryza sativa (C.-Q. Zhang et al., 2011) , Cc-nbs-llr, associated with pathogen recognition in plants (Meyers, Kozik, Griego, Kuang, & Michelmore, 2003), Necrotic spotted lesions 1, associated with negative regulation of SA-mediated defense response (Noutoshi et al., 2006), Subtilisin-like protease, associated with seed coat development (Tanaka et al., 2001), AP2-like ethylene-responsive transcription factor, associated with regulation of seed germination and response to stress (Licausi, OhmeâĂŘTakagi, & Perata, 2013), Aquaporin 1, involved in water channel transport (Murata, Mitsuoka, Hirai, & Walz, 2000), CBL-interacting protein kinase 9, associated with calcium-dependent signal transduction (Kolukisaoglu, Weinl, Blazevic, Batistic, & Kudla, 2004), and BEL1-like homeodomain protein 1, associated with response to abscisic acid (Dachan Kim et al., 2013).

3.3.3 Significant Pathways Associated with the Transcriptional Modifications

We also looked into the signal transduction pathways in which the differentially expressed genes under different stress conditions are involved. Studies have revealed that pathway-based analysis of gene expression data provides more reliable and in-

terpretable results as compared to gene expression data. Overall, a number of genes involved were found to be associated with 'Plant hormone signal transduction' and 'Biosynthesis of amino acids' pathways.

3.3.3.1 Pathways enriched due to psyllid herbivory

We were able to map 1359 of the genes that were differentially expressed due to herbivory by B. cockerelli under sufficient water (SPC) conditions to 27 different metabolic pathways belonging to the S. lycopersicum species using Over Representation Analysis. The most highly enriched pathways are 'Photosynthesis' with 8 differentially regulated genes, 'Carbon fixation in photosynthetic organisms' with 11 genes, and 'Nitrogen metabolism' with 5 genes. A list of top 10 metabolic pathways possibly regulated by B. cokerelli herbivory is presented in Fig. 3.3. The tables also include the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway identifier for each enriched pathway. Among plants affected by psyllid herbivory, plants under insufficient water conditions (SPD) showed enrichment in 31 different metabolic pathways. The highest enriched pathways are 'Phenylpropanoid biosynthesis' with 12 differentially regulated genes, 'Arginine biosynthesis' with 6 genes, and 'Cutin, suberine and wax biosynthesis' with 4 genes.

3.3.3.2 Pathways enriched due to Lso infection post psyllid herbivory

Plants under sufficient water conditions that were infected by Lso post pysllid herbivory (SLC) revealed enrichment in 26 different meatbolic pathways. The highest enriched pathways are 'Glutathione metabolism' with 11 genes, 'Cutin, suberine and wax biosynthesis' with 5 genes, 'Linoleic acid metabolism' with 3 genes and 'Pentose and glucuronate interconversions' with 6 genes. Among tomato plants affected by Lso infection post psyllid herbivory, plants under water shortage (SLD) showed enrichment in 24 metabolic pathways. The highly enriched pathways are 'Linoleic acid

metabolism', 'Biotin metabolism', 'Flavonoid Biosynthesis' and 'Zeatin biosynthesis'.

3.3.3.3 GO terms enriched

The transcripts were also subjected to gene ontology (GO) analysis. Under conditions of psyllid herbivory (SPC vs SCC), some of the GO terms enriched were lyase activity (GO:0016829), oxygen binding (GO: 0019825), ATPase activity (GO:0042626), glutathione transferase activity (GO:0004364), Response to biotic stimulus (GO: 0009607), structural constituent of cell wall (GO:0005199), auxin efflux transmembrane transporter activity (GO:0010329), cell wall (GO:0005618), DNA binding (GO:0003677). Under conditions of psyllid herbivory combined with drought (SPD vs SPC), GO terms were found to be associated with Response to biotic stimulus (GO: 0009607), Amino acid transmembrane transporter activity (GO:0015171), transmembrane receptor protein serine/threonine kinase activity (GO:0004675), oxygen binding (GO:0019825) and glutathione transferase activity (GO:0004364).

Under conditions of bacterial infestation and psyllid herbivory (SLC vs SCC), GO terms enriched were lyase activity (GO:0016829), lipid binding(GO:0008289), glutathione transferase activity (GO:0004364), oxygen binding (GO: 0019825), carbohydrate binding (GO:0030246). Under conditions of bacterial infestation and psyllid herbivory combined with drought (SLD vs SLC), GO terms were found to be associated with transferase activity (GO:0016740), low-affinity nitrate transmembrane transporter acitivity (GO:0080054), Carbohydrate binding (GO:0030246), cell wall (GO:0005618), Transcription factor activity (GO:0003700), oxygen binding (GO: 0019825), transmembrane receptor protein serine/threonine kinase activity (GO:0004675). The top 10 enriched GO terms under each of the stress conditions is provided in Fig. 3.4.

3.4 Discussion

Though plants are immobile, they have developed innate defense mechanisms to protect themselves from harmful interactions with herbivores, parasites, and adverse environmental conditions such as heat, drought, salinity, etc. In fact, studies suggest that the efficiency of this innate resistance to stress can be improved by subjecting the plant to milder stress conditions before exposing the plant to acute stress conditions. For instance, it was observed that plants exposed to cold and drought stress exhibited stronger resistance when they had been previously subjected to the same kind of stresses, a phenomenon referred to as acclimation (Hussain, 2011; Ramegowda et al., 2013).

Prior exposure to biotic stresses was also seen to result in stronger defense response when the plants were subsequently exposed to pathogens, (Conrath et al., 2006; Conrath, Beckers, Langenbach, & Jaskiewicz, 2015). Prior infection due to pathogens or treatments with other compounds has been shown to improve defense against subsequent pathogen attack by potentiating SA defense signaling (Kohler, Schwindling, & Conrath, 2002).

Inferences from the studies on acclimation in plants (Conrath, et al., 2006; Conrath, et al., 2015) suggest that the interaction between biotic and abiotic stresses may lead to enhancing resistance in some cases and reduced resistance in other cases. Drought acclimation in tomato has been reported to enhance resistance to the fungus Botrytis cinerea (Ramegowda, et al., 2013), whereas exposure to some abiotic stresses were found to have negative impacts on tomato plant defense response to pathogens (Ramegowda & Senthil-Kumar, 2015). In any case, the response of tomato plants to pathogen infection after exposure to abiotic stress conditions is not completely understood.

Research on plant response to combined stress situations (Achuo, Prinsen, & Hofte, 2006; McElrone & Forseth, 2004; Olson, Pataky, D'arcy, & Ford, 1990; Xu & Zhou, 2008) suggests that the result of the interactions depend upon the severity of each stress. For instance, plants subjected to mild drought stress conditions activate the basal defense response which facilitates them to show improved resistance to pathogen infection, whereas acute conditions of water shortage cause depletion of cellular nutrients leading to diminished protection against pathogens (Achuo, Prinsen, & Hofte, 2006). One of the purported reasons for stress tolerance due to acclimation or priming is the inherent capacity of plants to tailor response mechanisms based on stress specificity. Contrarily, plants could become susceptible to certain stresses due to the potential aggravation of damage caused by one stress.

The main goals of our study were to characterize transcriptomic changes of S. lycopersicum in response to phloem-feeding psyllids, and infection by the bacteria Lso, and to investigate how these responses are modified when the plants had been previously exposed to conditions of water shortage. Transcriptomic analysis of RNA sequences of samples under the different experimental conditions resulted in the detection of a large number of genes differentially regulated under these conditions, as shown in Fig. 3.1.

3.4.1 Psyllid Herbivory Leads to Widespread Transcriptomic Changes in Tomato

Herbivory of the phleom-feeding psyllid *B. cockerelli* led to extensive reprogramming in the transcriptome. This is seen in the large number of differentially expressed genes obtained in the psyllid herbivory cases as compared to just the drought conditions. In each case, the control samples were compared to the corresponding treatment samples. Genes identified in our study include transcription factors such as MYBs, BHLH, BZIP etc. and calcium sensors, kinases, calmodulin-binding chaper-

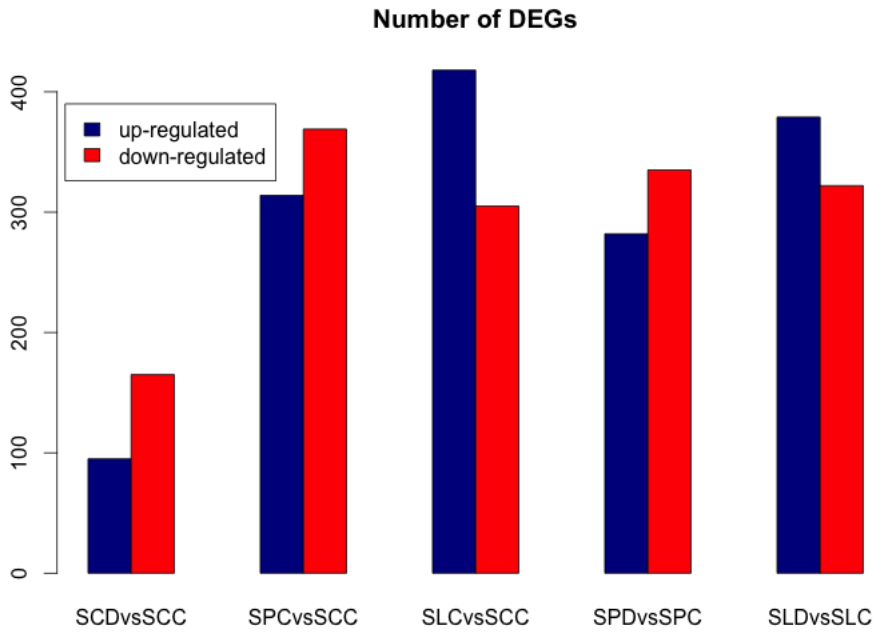


FIG. 3.1. Number of differentially expressed genes under the different experimental conditions.

onins, etc., many of which have previously been implicated, in defense response to biotic stresses including insect herbivory. The top 15 genes differentially expressed under psllid herbivory along with the log fold change and their functional description are tabulated in Fig. 3.5.

3.4.2 The Defense Response Elicited by Lso+psyllid Infection Has Overlaps with Response to Psyllid Herbivory

Lso is known to cause Zebra chip, a serious disease in potato and tomato crops (Julien Levy, Aravind Ravindran, Dennis Gross, Cecilia Tamborindeguy, & Elizabeth Pierson, 2011; Munyaneza, 2012). Plant responses to the combined stress caused by B. cockerelli herbivory and Lso are complex and the data presented here suggest that

the combined stress condition leads to regulation of multiple genes associated with plant responses to both psyllid herbivory and bacterial infection.

We found a large number of genes that were similarly regulated in the samples subjected to psyllid herbivory and in samples infected with Lso infection post psyllid herbivory. Fig. 3.2 is an illustration of the overlaps between differentially regulated genes under the two stress situations. Fig. 3.2 a) shows the overlap among up-regulated genes, while 2 b) shows the overlap among down-regulated genes.

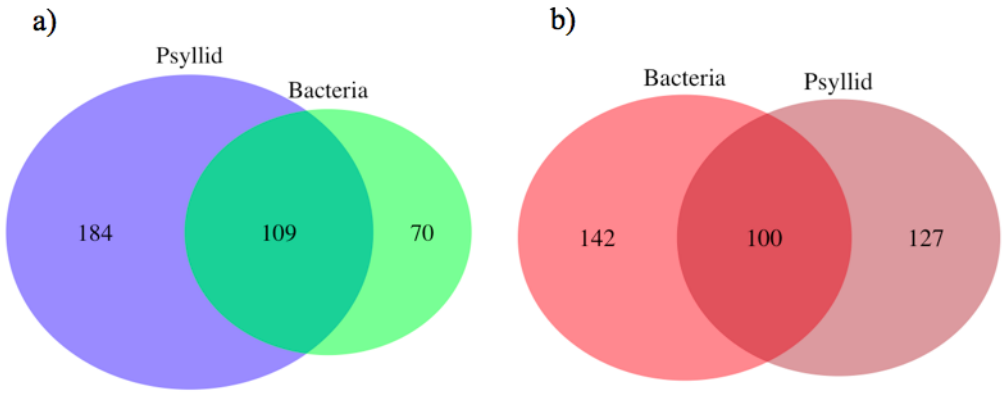


FIG. 3.2. Overlap among genes differentially expressed under psyllid herbivory and Lso infestation post psyllid herbivory. a) Overlap among up-regulated genes. b) Overlap among down-regulated genes.

Despite the similarities, some of the plant defenses regulated by JA (Jasmonic Acid) and SA (Salicylic Acid) seem to be suppressed under Lso infection. Certain known defense genes such as PR1, JAZ1, JAZ3 were significantly down-regulated in the Lso infected samples, but not in samples infected by psyllid herbivory alone.

Collectively, these data indicate that plants exhibit tailored responses to biotic stress stimuli and a number of these responses have overlapping components. How-

ever, the presence of drought leads to excessive complications in the responses of the tomato plant to different stresses, as discussed in the following paragraphs.

3.4.3 Plant Responses to Pathogen Infection + Drought Indicate Cross-talk between Pathways

Drought is known to be a major detrimental factor that inhibits plant growth and development. Though the observed responses of S. lycopersicum to combination of drought and psyllid herbivory (SPD vs SPC) had some overlap with the response to drought stress alone (SCD vs SCC), the major quantity of the genes differentially expressed in SPD vs SPC are unique to the combined stress situation.

Both the comparisons between SPD vs SPC and SCD vs SCC expressed significant up-regulation of HD-ZIP (Homeobox leucine zipper protein), HSP (Heat shock protein) and MYB transcripts, all of which are involved in plant defense response to abiotic stress stimuli (Sreenivasulu, Sopory, & Kishor, 2007). However, a large collection of genes differentially regulated in SPD vs SPC are absent in genes regulated by drought alone. In particular, a number of transcripts belonging to ABC transporter family and WRKY transcription factors are significantly up-regulated, uniquely in the combined stress condition. Transcripts related to Cold acclimation WCOR, Disease resistance response, Calcium/proton exchange, were also found to be characteristically up regulated in the combined stress condition. Similar behavior was observed for samples infected by Lso infection post psyllid herbivory when compared to drought samples.

Hence, our analyses indicate that conditions of water shortage in combination with other biotic stresses lead to extensive transformation in the transcriptome of tomato plants, unique to the combined stress situation.

3.5 Conclusions

Changing climatic conditions and the ever-growing population are affecting countries that depend on agriculture. Biotic and abiotic stress conditions lead to decrease in crop yields (Suzuki, et al., 2014). Amongst abiotic factors, water shortage affects the plant's capacity to produce food grains (Gallagher, Biscoe, & Hunter, 1976). Biotic stress factors such as insect herbivory and bacterial and fungal infestation lead to plant diseases and lowering quality of food crops (Oerke, 2006).

It has been observed that the characteristics of the applied stress determine the plant response to a large extent. In order to combat these harmful effects, it is imperative that we engineer plants with improved resistance. Identifying molecular factors that can confer desirable qualities is therefore essential. Comparative analysis of plant transcriptomes under different stress conditions is a highly useful tool in understanding plant responses to stress at the molecular level.

The main goal of this study was to get an understanding of the transcriptomic changes occurring in Solanum lycopersicum when subjected to psyllid herbivory of Bactericera cockerelli and bacterial infection of Lso under two different watering regimes. We identified differentially expressed transcripts under combined stress conditions and under single stress conditions.

We observed that plant responses to combined stress situations are vastly different from stress responses to individual stresses even though stresses of the same type were applied. Thus, this study underlines the importance of studying responses under combined stress conditions. Stress responsive genes that were discovered from this study will be useful in acquiring more cognition about drought stress responses in tomato plants. The transcripts identified from combined stress conditions of drought and Lso infection post psyllid herbivory can be used to engineer tolerance in plants

against these harmful pathogens in the presence of water shortage. Future experiments will be directed at carrying out combinatorial mutant studies involving some of the genes identified from the differential gene expression studies here.

KEGG Identifier	Pathway	Number of genes	Significant genes	P.VALUE
sly00196	Photosynthesis	18	8	8.23E-08
sly00710	Carbon_fixation_in_photosynthetic_organisms	53	11	1.25E-05
sly00910	Nitrogen_metabolism	12	5	1.41E-05
sly00904	Diterpenoid_biosynthesis	8	4	1.91E-05
sly01230	Biosynthesis_of_amino_acids	167	19	0.00036
sly00860	Porphyrin_and_chlorophyll_metabolism	38	7	0.000639
sly01200	Carbon_metabolism	180	19	0.000966
sly00040	Pentose_and_glucuronate_interconversions	32	6	0.00111
sly00950	Isoquinoline_alkaloid_biosynthesis	10	3	0.00123
sly02010	ABC_transporters	5	2	0.00134
sly00073	Cutin,_suberine_and_wax_biosynthesis	17	4	0.00143
sly00010	Glycolysis_/_Gluconeogenesis	86	11	0.00157
sly00030	Pentose_phosphate_pathway	36	6	0.00229
sly00630	Glyoxylate_and_dicarboxylate_metabolism	48	7	0.00314
sly00450	Selenocompound_metabolism	14	3	0.00496

FIG. 3.3. Putative Metabolic pathways involved in plant response to psyllid herbivory.

	GO Term	Number of genes	Significant genes	p-value
SCD vs SCC				
1	GO:0008559	11	6	5.71E-05
2	GO:0023034	9	5	0.000142
3	GO:0009607	13	6	0.000241
4	GO:0006869	14	6	0.000434
5	GO:0016829	11	5	0.000636
6	GO:0004702	5	3	0.000796
7	GO:0009922	13	5	0.00193
8	GO:0015381	6	3	0.00217
9	GO:0030246	18	6	0.00266
10	GO:0004675	78	17	0.00274
SLC vs SCC				
1	GO:0016829	11	5	2.08E-06
2	GO:0008289	13	5	7.20E-06
3	GO:0004364	19	6	7.24E-06
4	GO:0019825	69	9	0.000573
5	GO:0010329	5	2	0.000702
6	GO:0019901	5	2	0.000702
7	GO:0031683	5	2	0.000702
8	GO:0008559	11	3	0.000815
9	GO:0005618	12	3	0.00118
10	GO:0042626	12	3	0.00118
SPC vs SCC				
1	GO:0016829	11	6	3.42E-07
2	GO:0019825	70	13	1.64E-05
3	GO:0042626	12	4	0.000257
4	GO:0004364	19	5	0.000352
5	GO:0009607	13	4	0.000399
6	GO:0008289	14	4	0.000593
7	GO:0005507	23	5	0.00109
8	GO:0005199	5	2	0.00144
9	GO:0010329	5	2	0.00144
10	GO:0019901	5	2	0.00144

FIG. 3.4. GO terms enriched under different stress conditions.

SPC vs SCC - up regulated		
Gene	**Annotation**	**Log(Fold Change)**
Solyc01g109180.2.1	Long-chain-fatty-acid-CoA ligase	2.770405809
Solyc12g087940.1.1	Aspartic proteinase nepenthesin-1	2.552857127
Solyc01g111880.2.1	Tyrosine-protein kinase transforming protein Src	2.451809683
Solyc04g005660.2.1	Transcription factor style2.1	2.363569863
Solyc02g086820.2.1	Carbonic anhydrase	2.310576201
Solyc06g065970.1.1	Cortical cell-delineating protein	2.177009838
Solyc03g025720.2.1	Long-chain-fatty-acid--CoA ligase	2.134658074
Solyc07g062710.2.1	BZIP transcription factor family protein	2.077927535
Solyc05g009890.1.1	Chloroplast nucleoid DNA binding protein-like	2.067368366
Solyc07g055950.2.1	Meiosis 5	1.969657708
Solyc05g014000.2.1	Pectate lyase	1.964350467
Solyc01g090970.2.1	Cortical cell-delineating protein	1.890657431
Solyc07g042560.2.1	Kinesin-like protein	1.885196234
Solyc08g080660.1.1	Osmotin-like protein (Fragment)	1.847761203
Solyc02g070180.1.1	FAD-binding domain-containing protein	1.808860438

SPC vs SCC - down regulated		
Gene	**Annotation**	**Log(Fold Change)**
Solyc08g075570.2.1	Urea active transporter-like protein	-3.503529688
Solyc09g011520.2.1	Glutathione S-transferase-like protein	-3.270256007
Solyc05g009550.2.1	IST1 homolog	-3.145646631
Solyc03g098740.1.1	Kunitz trypsin inhibitor	-3.051001428
Solyc07g005100.2.1	Chitinase-like protein	-2.819424034
Solyc01g097240.2.1	Pathogenesis-related protein 4B (Fragment)	-2.661063627
Solyc06g072350.2.1	UPF0497 membrane protein 17	-2.599123373
Solyc09g091670.2.1	ATP-binding cassette transporter	-2.464403286
Solyc07g048070.2.1	Membrane protein	-2.326890203
Solyc02g079510.2.1	Peroxidase	-2.266464796
Solyc09g082760.2.1	Aspartic proteinase 2	-2.193032924
Solyc09g091000.2.1	Major allergen Mal d 1	-2.182501833
Solyc02g084410.2.1	Glyoxalase	-2.160575584
Solyc03g116700.2.1	Blue copper protein	-2.139709981
Solyc07g008020.2.1	Auxin response factor 16	-2.095169019

FIG. 3.5. Significant genes differentially expressed under psyllid herbivory along with their functional description and log_2 fold change values.

4. ANALYSIS OF THE EFFECT OF METFORMIN ON CELLS USING HIGH-CONTENT EPIFLUORESCENT IMAGING DATA*

4.1 Introduction

Metformin, an FDA-approved biguanide, is known to reduce levels of circulating glucose and is widely used for the treatment of diabetes mellitus. Epidemiological studies of cancer patients with diabetes [124] led to the initial indication that metformin may be associated with reduced cancer-related mortality. Subsequently, a number of in vitro and in vivo studies on cancer cells have reported a negative association between cancer risk and the use of metformin [125], [126]. As a result, there has been an increasing effort to understand the mechanism of action of metformin in an attempt to reposition the drug for the treatment of cancer.

The effects of metformin on diabetic patients diagnosed with colorectal cancer was studied in [127] by means of survival analysis. Data pertaining to 595 patients diagnosed with both colorectal cancer and diabetes mellitus were analyzed, and metformin use was found to be associated with decreased colorectal cancer specific mortality. A population-based cohort study in [128] also connected metformin use with reduced colon cancer risk. Consequently, metformin is being debated for use as a prophylactic agent for colon cancer.

The anti-cancer effect of metformin has in part been attributed to its up-regulation of Adenosine Monophosphate (AMP) activated protein kinase activity [129], [130]. Studies have suggested that the activation of AMPK by metformin further leads to

*Parts of this section are reprinted with permission from "Epifluorescent imaging study of the effect of anti-diabetic drug metformin on colorectal cancer cell lines in vitro" by Venkatasubramani P, Sima C, Hua J, Cypert M, Bittner M, Datta A, 2017. *Journal of Cancer Research & Therapy*, volume 5, no. 4, pages 19 - 23, ISSN 2052-4994. http://dx.doi.org/10.14312/2052-4994.2017-4.

inhibition of mammalian Target of Rapamycin (mTOR) in breast cancer cells [131].

A study on the effect of metformin in breast cancer cell lines reported dose-dependent reduction in cell proliferation in cell lines MCF-7 and MCF-10A [130]. As per this study, metformin in concentrations of 2.5-20 mM led to a reduction of about 40-70% in cellular proliferation of cancer cells after exposure for 72 hours. (Cell proliferation measurement in [130] was carried out by calculating percentage reduction in the intensity of Alamar Blue dye at the end of the experiment.)

A follow-up of the study conducted in [130] reported that the use of metformin led to reduction in global protein synthesis and also decreased cap-dependent translation in Mouse Embryonic Fibroblasts (MEFs) [131]. The major cause of these effects of metformin was reported to be its inhibition of mTOR. Exposure to Metformin for 72 hours in doses of 20 mM were reported to inhibit the growth of TSC2+/+ (tuberous sclerosis complex 2) MEFs by 53% as compared to the control samples. (MEFs were stained with crystal violet, which was eluted with acetic acid, and light absorption levels of the supernatant was measured for each sample at 570 nm to determine growth in cells.)

Many such studies correlated the use of metformin with better prognosis for cancer patients because of the observed reduction in cellular proliferation. While a number of experiments performed in these studies [130], [131] showed a reduction in the total number of cells on treatment with metformin, there is a lot of variance in the amount of reduction observed. Furthermore, it is not clear if the use of metformin killed cancer cells or retarded the growth of cells in the cancerous cell lines, since the conventional cell death assays only provide snapshots of cellular proliferation at the initial and final time points of the experiment.

In order to capture the full dynamics of the system over an extended period of time, multiple assays need to be produced under identical conditions at different time

points by extracting analytes at each time point from a different cell population. In order to circumvent this cumbersome procedure, a transcriptional fluorescent imaging technique is utilized to directly follow the death of a set of cells over time in the current study.

Fluorescent reporters have long been used for temporal studies in experimental biology in order to study transcriptional activities of cells. In this study, a stream of images of a particular population of cells is acquired from the beginning to the end of the experiment at different closely spaced time points. These images are processed to obtain time course data that contains details about cellular activities during the course of the experiment. This data is processed and visualized in order to attain a deeper understanding of the dynamics at the cellular level.

To track cell death caused by treatment with metformin, colon cancer cells were stained with CellTox green dye (details explained in the Materials and Methods section), which preferentially stains dead cell DNA when cell membrane integrity is compromised, producing a nucleus specific, bright green fluorescent signal. The fluorescent signals generated in the experiment were tracked by repeatedly capturing images of the same sites and then scoring cell life and death so as to quantify cell death at different time points over a period of 37 hours.

4.2 Materials and Methods

4.2.1 High-content Epifluorescent Imaging

A single assay in our experimental setup consists of epifluorescent imaging of a spot at the bottom in a 384-well plate, thus generating an image of the cells in that region (200 to 300 cells) bearing fluorescent reporters. ImageXpress Micro XLS High-Content Imaging System (Molecular Devices, LLC) was used to capture the time lapse fluorescent image data from cells in multi-well plates. For studying cell

death, fluorescent images are extracted as two-color image sets with a blue channel image for the nuclei and a green channel image for the fluorescent CellTox Green reporter. When cell death occurs and the cell membrane collapses, the CellTox Green dye stains the dead cell DNA with fluorescent green color.

4.2.2 Quantification of Cell Death from Images

Visual examination of images may provide some indication of the extent of cell death at a particular imaging site, but advanced image processing procedures followed by efficient data summarization help in acquiring reliable estimates of cell death occurring throughout the experiment. In order to assess the amount of cell death caused by metformin, certain morphological features of the cell are used to label a cell as dead or alive. Primary among these features is disruption of the cell membrane, at which time the CellTox Green dye enters the cell and binds to the DNA. Additionally, pyknosis, (condensation of chromatin in the nucleus) which can be detected on the nuclear channel of the images, is another key feature of cell death. The image-processing pipeline detailed in [132] establishes thresholds for nuclear size and intensity of the dye in the nucleus in order to detect cell membrane collapse, and these thresholds are used to count the number of dead cells in the images at various sites in the well. The percentage cell death is then calculated as: % Cell death = number of dead cells / total number of cells x 100.

4.2.3 Statistical Analysis

Percentage cell death at each imaging site for the different doses of metformin were tested for statistical significance. Statistical analysis was performed using the t-test with R software, version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria) [133]. $P<0.05$ was considered to indicate a statistically significant difference.

4.2.4 Cell Lines and Treatments

Cell lines HCT116 and SW480 were plated at a density of 6000 cells/well in 30 μl/well of Imaging Media (IM) on one 384-well microtiter plate (Greiner Bio-One 781 09x) pre-coated with 10 g/ml Rat Tail Collagen Type I (BD Biosciences 354249) and washed 3X with sterile 1X PBS without calcium and magnesium. IM consists of 70% M-199 (Thermo Fisher Scientific, 11825015), 30% RPMI-1640 (11875085) supplemented with 10% FBS (16000044), 11 mM D-glucose (A2494001), 20 mM HEPES (15630080) and 20 mM GlutaMax (35050061). Nuclei of the cells were stained with Vybrant DyeCycleViolet live-cell nuclear stain (ThermoFisher Scientific) diluted 1:15,000. The cells were allowed to attach to the surface of the plate by incubating at 37°C with 5% CO2 for 6 hours.

After taking images of cells at the first baseline time point before treatment, the cells were treated with 5 mM, 10 mM, 15 mM, and 20 mM concentrations of Metformin diluted in IM or left untreated as control. Images of each well were taken every hour up to 36 hours using the High-Content Imaging System.

4.3 Materials and Methods

4.3.1 High-content Epifluorescent Imaging

ImageXpress Micro XLS High-Content Imaging System (Molecular Devices, LLC) was used to capture time lapse fluorescent image data from cells in multi-well plates. Cell-level dynamics extracted from images of wells were used to quantify live and dead cells using a combination of advanced image processing techniques and data representation methods [132].

A fluorescent protein based promoter-reporter technology was adapted for monitoring gene expression patterns for a set of genes. Whenever the gene of interest is transcribed, the reporter will also be transcribed (and translated), so that the

expression level of the gene of interest can be detected by measuring the intensity of fluorescence of the fluorescent protein.

This state-of the-art imaging technology was used to follow changes in mRNA promoter transcriptional activity of multiple reporters in parallel. The samples were taken at 3 different imaging sites within each well. The relative transcriptional activity of genes is illustrated using bar plots, which facilitates summarized representation of multiple gene activities over a long period of time.

4.3.2 Cell Lines and Treatments

Cell lines HCT116 and SW480 permanently expressing fluorescent promoter reporters IL6 eGFP, NFKB1 eGFP and MAP1LC3A eGFP were previously created using a lentiviral expression system based on pLenti6/V5-DEST Gateway Vector platform (ThermoFisher Scientific, V496-10). Cells of each type were plated at a density of 6000 cells/well in 30 l/well of Imaging Media (IM) on one 384-well microtiter plate (Greiner Bio-One 781 09x) pre-coated with 10 μg/ml Rat Tail Collagen Type I (BD Biosciences 354249) and washed 3X with sterile 1X PBS without calcium andÂămagnesium. Nuclei of the cells were stained with Vybrant ®DyeCycle Violet live-cell nuclear stain (ThermoFisher Scientific) diluted 1:15,000. The cells were allowed to attach to the surface of the plate by incubating at 37°C with 5% CO2 for 6 hours. After taking images of cells at the first baseline time point before treatment, the cells were treated with 5 mM, 10 mM, 15 mM and 20 mM concentrations of Metformin diluted in IM or left untreated as control. Images of each well where taken every hour up to 36 hours using the High-Content Imaging System.

4.4 Results

4.4.1 Metformin Does Not Cause Significant Cell Death in Colon Cancer Cell Lines In Vitro

We examined the effect of metformin on cell death in colon cancer cell lines HCT116 and SW480 through high content epifluorescent imaging. Fig. 4.1 shows a continuous summarized representation of percentage cell death measured on exposure of colon cancer cell lines to metformin over a period of 37 hours. Met5, Met10, Met15 and Met20 correspond to 5mM, 10mM, 15mM and 20mM doses of metformin, respectively. The percentage cell death values shown are means of cell death values calculated at different imaging sites for each treatment condition. As seen from Fig. 4.1 a), in cell line HCT116, an increase of about 20% in cell death was observed on average at certain time points on exposure to metformin, but continuous increase was not sustained through the experiment. Fig. 4.1 b) shows that relative to the control, metformin caused no significant increase in cell death in the cell line SW480. Fig. 4.2 shows the cell death percentage at the different doses and at 4 different time points (24h, 28h, 32h, 36h) in the HCT116 cell line, relative to the control. Cells were treated with 5-20 mM metformin for 37 hours, and percentage cell death was quantified as outlined in the Materials and Methods. Results are given as means $\pm$ Standard Error for 9 replicate determinations at 4 different imaging sites for each treatment, and significant ($p < 0.05$) increase in cell death is indicated (*). Fig. 4.3 shows the cell death in the SW480 cell line at 4 different time points, relative to the control. Significant ($p < 0.05$) increase in cell death is indicated (*).

4.5 Discussion

Drug repositioning, a term used to refer to the identification of new applications for existing drugs, has become popular in recent years, fueled by the large scale

battle against cancer and the search for anti-cancer drugs by the world-wide cancer research community. Existing drugs have already been in human use and have minimal cell toxicity, so they are appealing alternatives to discovering new drugs that might require extensive study and testing before they can be presented for FDA approval.

A potential link between metabolism and cancer has led to multiple investigations into the effects of metformin on different cancers with the ultimate goal of repositioning metformin to treat cancers. Our investigations into the action of metformin, aimed at studying the cell death caused by metformin in cancer cell lines, reveal that high concentrations of metformin caused some cell death in the cell line HCT116, which was significant up to 32 h, while there was no significant cell death observed in the cell line SW480. Such inconsistent effect of metformin on cell lines has been reported before in [131], where the authors observed dose-dependent effect of metformin on breast and ovarian cancer cell lines, and no effect on the HeLa cell line. This indicates that not all cancer cell lines respond in a similar way to treatment with metformin, and individual studies on different cell lines are necessary.

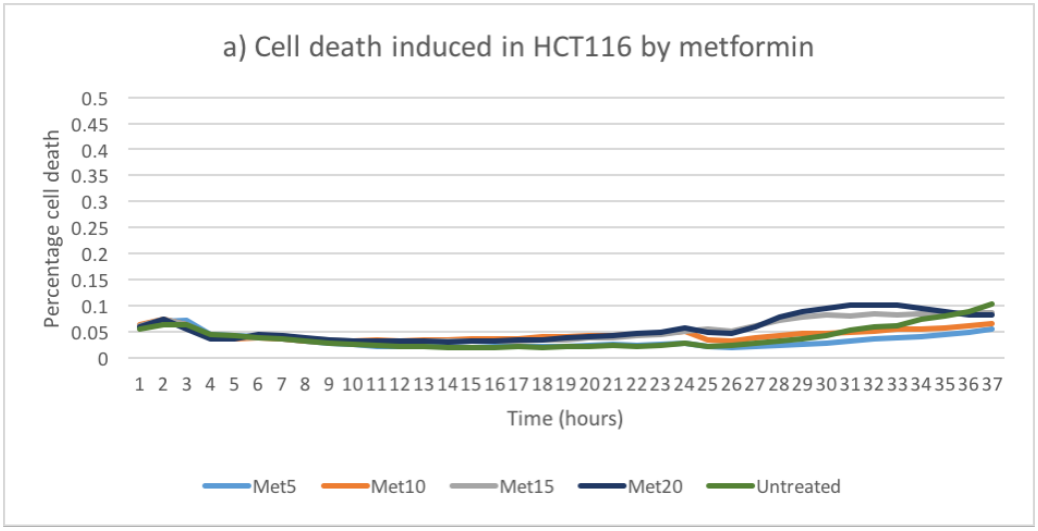
In our study, the colon cancer cell lines were monitored up to 37 hours. It is possible that there may be more cell death observed if the cells are tracked for longer time. However, the trends we observe make such a possibility unlikely. Moreover, rapid cell killing is a highly desired characteristic in a chemotherapy drug. Based on our study, metformin, when administered individually to colon cancer cell lines, does not seem to possess this characteristic.

The experiments carried out in our study are in vitro, where the cells are treated outside their natural environment. Consequently, intracellular signaling in the cells under study is impaired, and the responses observed may not truly represent responses from an in vivo animal study or a clinical trial. The inherent abnormality

and heterogeneity of cancer tissues also make it very challenging to relate outcomes from in vitro models to human cancer outcomes. However, in vitro methods are widely established, and facilitate ease of interpretation of results since the composition of the cells being analyzed is well known beforehand. In vitro studies are also good indicators of risk and underlying biological mechanisms, and must be analyzed meticulously before proceeding to in vivo animal studies or clinical trials. The lack of clear patterns in the amount of cell death observed at different doses and time points and the large variability in the results for different cell lines in our study indicate that we need much more information about how metformin affects different cancer cell lines. It is imperative that this deficiency in understanding the effect of metformin be kept in mind when considering novel therapies based on metformin.

Interestingly, metformin in combination with doxorubicin, a chemotherapeutic drug, was found to inhibit the growth of cancer cells in culture in four different breast cancer cells MCF-7, MCF-10A, SKBR3 and MDA-MB-486 [134]. Additionally, in this study, MCF-10A cells were injected into nude mice and the tumor volume was measured at various time intervals, and it was found that the combination therapy reduced the tumor volume and prevented relapse of the cancer more effectively as compared to either drug alone. The main reason for this observation was reported to be the ability of metformin to selectively kill cancer stem cells. Experimental studies in [135], [136] have also indicated that metformin may potentiate the effect of certain chemotherapeutic drugs such as cisplatin, carboplatin, etc., when used in combination in the treatment of cancer cells. A great advantage of combining a drug metformin in a chemotherapeutic cocktail with a more potent drug is that it may help in reducing the cytotoxicity of the overall dose, while maintaining the level of cell killing achieved by the standard chemotherapy drug. Hence, the utility of metformin may be in combination therapy for treating cancer, rather than individually. Our

future studies will be aimed at examining the action of metformin in combination with other chemotherapy drugs.



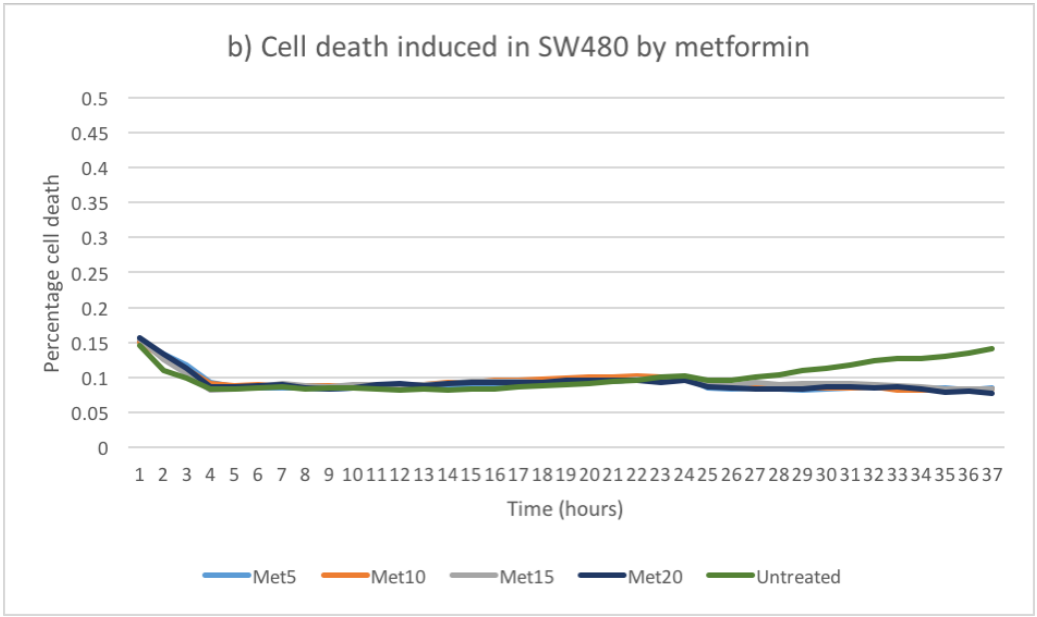


FIG. 4.1. Percentage cell death relative to the untreated cells (control) are shown at several time points after the administration of the drug metformin in different doses to colon cancer cell line a) HCT116 and b) SW480.

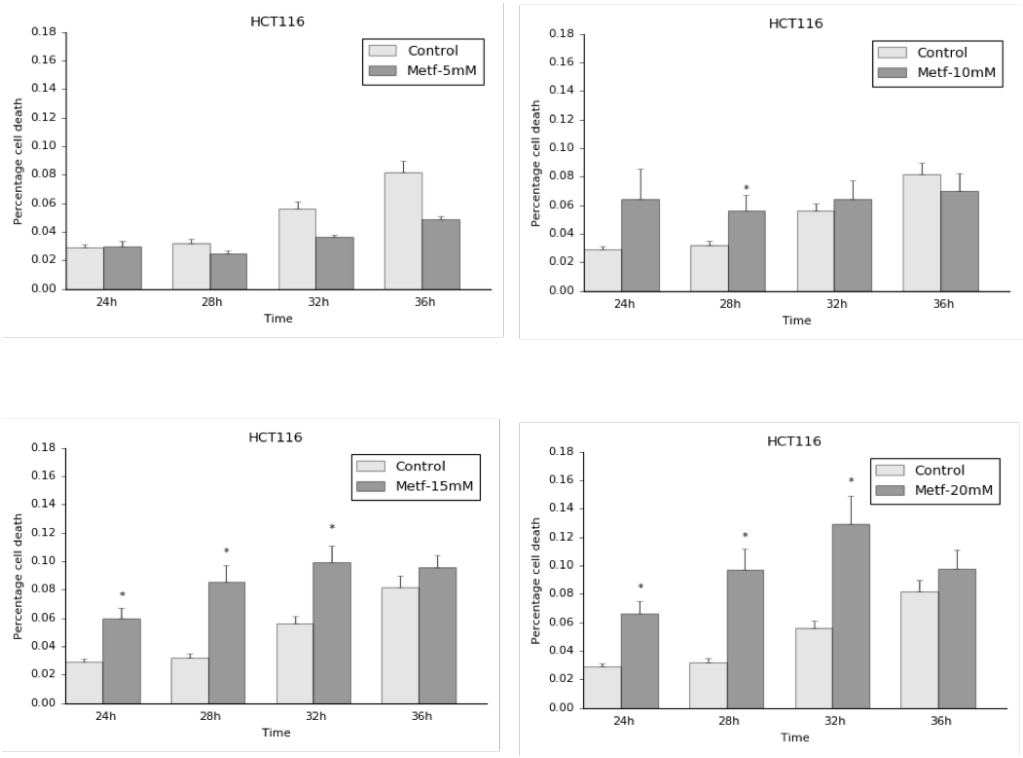


FIG. 4.2. Cell death induced by metformin in colon cancer cell line HCT116.

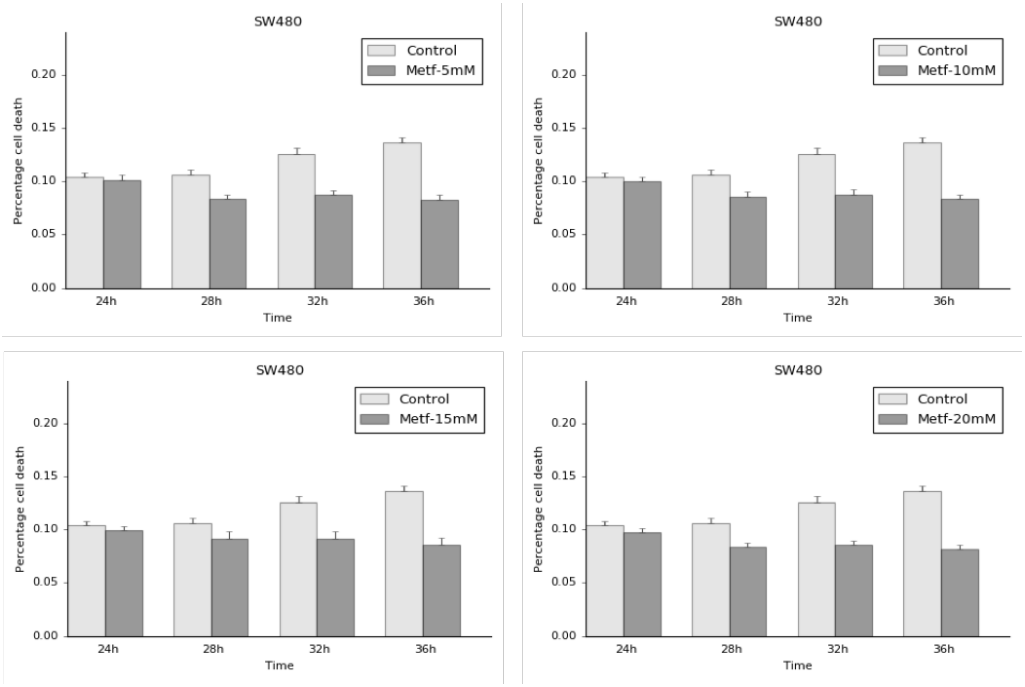


FIG. 4.3. Cell death induced by metformin in colon cancer cell line SW480.

5. CONCLUSIONS AND FUTURE WORK

In this dissertation we have shown how Bayesian learning and inference can help in making useful deductions by integrating with biological data. We also showed methods to tackle the challenges in understanding big data in genomics.

5.1 Summary of Conclusions

In Chapter 2, we described a general framework for quantifying the effect of an intervention in a gene regulatory network using Bayesian Network Models. We have demonstrated a method to fuse existing knowledge about gene interactions with actual experimental data, and also demonstrated a means to use this method to make decisions related to selection of genes for intervention.

In Chapter 3, we showed how advanced approaches to modeling count data can be used for understanding the response of plants to attack by insects and pathogens. In Chapter 4, we demonstrated how image processing methods and statistical techniques can be used to evaluate the utility of the diabetic drug Metformin in a combination therapy for treating cancer.

5.2 Future Work

We are currently working on improving methods for various types of genomic data analysis by expanding the Bayesian Network framework described in this work. Promising directions for such extensions include: (a) Using Dynamic Bayesian Network Models by making use of temporal gene expression or sequencing data; (b) Integrating data of more than one type into the same model in order to fully exploit all available information, and also facilitate making robust decisions from the model.

REFERENCES

[1] Strange, R.N. and Scott, P.R., 2005. Plant disease: a threat to global food security. Annu. Rev. Phytopathol., 43, pp.83-116.

[2] Brader, G., Djamei, A., Teige, M., Palva, E.T. and Hirt, H., 2007. The MAP kinase kinase MKK2 affects disease resistance in Arabidopsis. Molecular Plant-Microbe Interactions, 20(5), pp.589-596.

[3] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning, 46(1), pp.389-422.

[4] Shevade, S.K. and Keerthi, S.S., 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, 19(17), pp.2246-2253.

[5] Díaz-Uriarte, R. and De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), p.3.

[6] Ben-Dor, A., Shamir, R. and Yakhini, Z., 1999. Clustering gene expression patterns. Journal of computational biology, 6(3-4), pp.281-297.

[7] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. and Sampas, N., 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 406(6795), pp.536-540.

[8] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12), pp.6745-6750.

[9] Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of theoretical biology, 22(3), pp.437-467.

[10] Mestl, T., Plahte, E. and Omholt, S.W., 1995. A mathematical framework for describing and analysing gene regulatory networks. Journal of Theoretical Biology, 176(2), pp.291-300.

[11] Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W., 2002. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 18(2), pp.261-274.

[12] Friedman, N., Linial, M., Nachman, I. and Pe'er, D., 2000. Using Bayesian networks to analyze expression data. Journal of computational biology, 7(3-4), pp.601-620.

[13] Werhli, A.V. and Husmeier, D., 2007. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol, 6(1), p.15.

[14] Langfelder, P. and Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), p.559.

[15] Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. and Romero, R., 2009. A novel signaling pathway impact analysis. Bioinformatics, 25(1), pp.75-82.

[16] Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M., 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 26(12), pp.i237-i245.

[17] Zhu, Y., Xu, Y., Helseth, D.L., Gulukota, K., Yang, S., Pesce, L.L., Mitra, R., Müller, P., Sengupta, S., Guo, W. and Silverstein, J.C., 2015. Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data. Journal of the National Cancer Institute, 107(8), p.djv129.

[18] Pearl, J., 2014. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.

[19] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M., 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. science, 302(5644), pp.449-453.

[20] Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J. and d'Alche-Buc, F., 2003. Gene networks inference using dynamic Bayesian networks. Bioinformatics, 19(suppl 2), pp.ii138-ii148.

[21] Schäfer, J. and Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics, 21(6), pp.754-764.

[22] Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on systems, man, and cybernetics, 9(1), pp.62-66.

[23] Hoff, P.D., 2009. A first course in Bayesian statistical methods. Springer Science & Business Media.

[24] Vazquez, A., Bond, E.E., Levine, A.J. and Bond, G.L., 2008. The genetics of the p53 pathway, apoptosis and cancer therapy. Nature reviews Drug discovery, 7(12), pp.979-987.

[25] Aumann, R.J., 1995. Backward induction and common knowledge of rationality. Games and Economic Behavior, 8(1), pp.6-19.

[26] Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.

[27] de Wit, P.J., 2007. How plants recognize pathogens and defend themselves. Cellular and Molecular Life Sciences, 64(21), pp.2726-2732.

[28] Taiz, Lincoln, Eduardo Zeiger, Ian Max Møller, and Angus Murphy. Plant physiology and development. Sinauer Associates, Incorporated, 2015.

[29] Galán, J.E. and Collmer, A., 1999. Type III secretion machines: bacterial devices for protein delivery into host cells. Science, 284(5418), pp.1322-1328.

[30] Keen, N.T., 1982. Specific recognition in gene-for-gene host-parasite systems [Plant pathogens, phytoalexins]. Advances in plant pathology.

[31] Meng, X. and Zhang, S., 2013. MAPK cascades in plant disease resistance signaling. Annual Review of Phytopathology, 51, pp.245-266.

[32] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research, 42(D1), pp.D199-D205.

[33] Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), pp.27-30.

[34] Edgar, R., Domrachev, M. and Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), pp.207-210.

[35] Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. and Hornik, K., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), p.R80.

[36] Wang, Q., Wang, M., Zhang, X., et. al. 2011. WRKY gene family evolution in Arabidopsis thaliana. *Genetica.* 139, 973-983.

[37] Zhao, C., Nie, H., Shen, Q., et. al. 2014. EDR1 physically interacts with MKK4/MKK5 and negatively regulates a MAP kinase cascade to modulate plant innate immunity. *PLoS Genet.*, 10, e1004389.

[38] Achuo, E.A., Prinsen, E. and Höfte, M., 2006. Influence of drought, salt stress and abscisic acid on the resistance of tomato to Botrytis cinerea and Oidium neolycopersici. *Plant pathology*, 55(2), pp.178-186.

[39] Alves, M.S., Dadalto, S.P., Goncalves, A.B., De Souza, G.B., Barros, V.A. and Fietto, L.G., 2013. Plant bZIP transcription factors responsive to pathogens: a review. International journal of molecular sciences, 14(4), pp.7815-7828.

[40] An, C.I., Fukusaki, E.I. and Kobayashi, A., 2002. Aspartic proteinases are expressed in pitchers of the carnivorous plant Nepenthes alata Blanco. Planta, 214(5), pp.661-667.

[41] Anders, S. and Huber, W., 2010. Differential expression analysis for sequence count data. Genome biology, 11(10), p.R106.

[42] Atkinson, N.J. and Urwin, P.E., 2012. The interaction of plant biotic and abiotic stresses: from genes to the field. Journal of experimental botany, 63(10), pp.3523-3543.

[43] Bennett, S., 2004. Solexa ltd. Pharmacogenomics, 5(4), pp.433-438.

[44] Boari, F. and Malone, M., 1993. Rapid and systemic hydraulic signals are induced by localized wounding in a wide range of species. J Exp Bot, 44, pp.741-746.

[45] Bowles, D.J., 1990. Defense-related proteins in higher plants. Annual review of biochemistry, 59(1), pp.873-907.

[46] Brown, J.K., Rehman, M., Rogan, D., Martin, R.R. and Idris, A.M., 2010. First report of "Candidatus Liberibacter psyllaurous"(synonym "Ca. L. solanacearum") associated with 'tomato vein-greening' and 'tomato psyllid yellows' diseases in commercial greenhouses in Arizona. Plant Disease, 94(3), pp.376-376.

[47] Casteel, C.L., Hansen, A.K., Walling, L.L. and Paine, T.D., 2012. Manipulation of plant defense responses by the tomato psyllid (Bactericerca cockerelli) and its associated endosymbiont Candidatus Liberibacter psyllaurous. PloS one, 7(4), p.e35191.

[48] Chinnusamy, V., Schumaker, K. and Zhu, J.K., 2004. Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. Journal of experimental botany, 55(395), pp.225-236.

[49] Chung, H.S., Koo, A.J., Gao, X., Jayanty, S., Thines, B., Jones, A.D. and Howe, G.A., 2008. Regulation and function of Arabidopsis JASMONATE ZIM-domain genes in response to wounding and herbivory. Plant physiology, 146(3), pp.952-964.

[50] Conrath, U., Beckers, G.J., Flors, V., García-Agustín, P., Jakab, G., Mauch, F., Newman, M.A., Pieterse, C.M., Poinssot, B., Pozo, M.J. and Pugin, A., 2006. Priming: getting ready for battle. Molecular Plant-Microbe Interactions, 19(10), pp.1062-1071.

[51] Conrath, U., Beckers, G.J., Langenbach, C.J. and Jaskiewicz, M.R., 2015. Priming for enhanced defense. Annual review of phytopathology, 53, pp.97-119.

[52] Crosslin, J.M., Lin, H. and Munyaneza, J.E., 2011. Detection of 'Candidatus Liberibacter Solanacearum' in the Potato Psyllid, Bactericera cockerelli (Sulc) 1, by Conventional and Real-Time PCR. Southwestern Entomologist, 36(2), pp.125-135.

[53] Datta, K., Velazhahan, R., Oliva, N., Ona, I., Mew, T., Khush, G.S., Muthukrishnan, S. and Datta, S.K., 1999. Over-expression of the cloned rice thaumatin-like protein (PR-5) gene in transgenic rice plants enhances environmental friendly resistance to Rhizoctonia solani causing sheath blight disease. TAG Theoretical and Applied Genetics, 98(6), pp.1138-1145.

[54] Ding, X., Gopalakrishnan, B., Johnson, L.B., White, F.F., Wang, X., Morgan, T.D., Kramer, K.J. and Muthukrishnan, S., 1998. Insect resistance of transgenic tobacco expressing an insect chitinase gene. Transgenic research, 7(2), pp.77-84.

[55] Easlon, H.M. and Richards, J.H., 2009. Drought response in self-compatible species of tomato (Solanaceae). American Journal of Botany, 96(3), pp.605-611.

[56] Fan, W. and Dong, X., 2002. In vivo interaction between NPR1 and transcription factor TGA2 leads to salicylic acid-mediated gene activation in Arabidopsis. The Plant Cell, 14(6), pp.1377-1389.

[57] Farooq, M., Wahid, A., Kobayashi, N., Fujita, D. and Basra, S.M.A., 2009. Plant drought stress: effects, mechanisms and management. In Sustainable agriculture (pp. 153-188). Springer Netherlands.

[58] Fujimoto, S.Y., Ohta, M., Usui, A., Shinshi, H. and Ohme-Takagi, M., 2000. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. The Plant Cell, 12(3), pp.393-404.

[59] Gallagher, J.N., Biscoe, P.V. and Hunter, B., 1976. Effects of drought on grain growth. Nature, 264(5586), pp.541-542.

[60] Galon, Y., Nave, R., Boyce, J.M., Nachmias, D., Knight, M.R. and Fromm, H., 2008. CalmodulinâĂŘbinding transcription activator (CAMTA) 3 mediates biotic defense responses in Arabidopsis. FEBS letters, 582(6), pp.943-948.

[61] Geistlinger, L., Csaba, G. and Zimmer, R., 2016. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set-& network-based enrichment analysis. BMC bioinformatics, 17(1), p.45.

[62] Giunta, F., Motzo, R. and Deidda, M., 1993. Effect of drought on yield and yield components of durum wheat and triticale in a Mediterranean environment. Field Crops Research, 33(4), pp.399-409.

[63] Glazebrook, J., 2001. Genes controlling expression of defense responses in ArabidopsisâĂŤ2001 status. Current opinion in plant biology, 4(4), pp.301-308.

[64] Gong, P., Zhang, J., Li, H., Yang, C., Zhang, C., Zhang, X., Khurram, Z., Zhang, Y., Wang, T., Fei, Z. and Ye, Z., 2010. Transcriptional profiles of drought-responsive genes in modulating transcription signal transduction, and biochemical pathways in tomato. Journal of experimental botany, 61(13), pp.3563-3575.

[65] Hermans, C., Porco, S., Verbruggen, N. and Bush, D.R., 2010. Chitinase-like protein CTL1 plays a role in altering root system architecture in response to multiple environmental conditions. Plant Physiology, 152(2), pp.904-917.

[66] Holmstrom, K.O., Mantyla, E., Wellin, B. and Mandal, A., 1996. Drought tolerance in tobacco. Nature, 379(6567), p.683.

[67] S. Sanghera, G., H. Wani, S., Hussain, W. and B. Singh, N., 2011. Engineering cold stress tolerance in crop plants. Current genomics, 12(1), p.30.

[68] Igawa, T., Ochiai-Fukuda, T., Takahashi-Ando, N., Ohsato, S., Shibata, T., Yamaguchi, I. and Kimura, M., 2004. New TAXI-type xylanase inhibitor genes are inducible by pathogens and wounding in hexaploid wheat. Plant and cell physiology, 45(10), pp.1347-1360.

[69] Izawa, T., Foster, R. and Chua, N.H., 1993. Plant bZIP protein DNA binding specificity. Journal of molecular biology, 230(4), pp.1131-1144.

[70] Jin, L., Zuo, X.Y., Su, W.Y., Zhao, X.L., Yuan, M.Q., Han, L.Z., Zhao, X., Chen, Y.D. and Rao, S.Q., 2014. Pathway-based analysis tools for complex diseases: a review. Genomics, proteomics & bioinformatics, 12(5), pp.210-220.

[71] Jones, H.G., Flowers, T.J. and Jones, M.B., 1989. Plants under stress: biochemistry, physiology and ecology and their application to plant improvement (Vol. 39). Cambridge University Press.

[72] Khatri, P. and Drăghici, S., 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18), pp.3587-3595.

[73] Kim, D., Cho, Y.H., Ryu, H., Kim, Y., Kim, T.H. and Hwang, I., 2013. BLH1 and KNAT3 modulate ABA responses during germination and early seedling development in Arabidopsis. The Plant Journal, 75(5), pp.755-766.

[74] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology, 14(4), p.R36.

[75] Kodde, D.A. and Palm, F.C., 1986. Wald criteria for jointly testing equality and inequality restrictions. Econometrica: journal of the Econometric Society, pp.1243-1248.

[76] Kohler, A., Schwindling, S. and Conrath, U., 2002. Benzothiadiazole-induced priming for potentiated responses to pathogen infection, wounding, and infiltration of water into leaves requires the NPR1/NIM1 gene in Arabidopsis. Plant Physiology, 128(3), pp.1046-1056.

[77] Kolukisaoglu, U., Weinl, S., Blazevic, D., Batistic, O. and Kudla, J., 2004. Calcium sensors and their interacting protein kinases: genomics of the Arabidopsis and rice CBL-CIPK signaling networks. Plant physiology, 134(1), pp.43-58.

[78] Kunkel, Barbara N., and David M. Brooks. Cross talk between signaling pathways in pathogen defense. Current opinion in plant biology 5.4 (2002): 325-331.

[79] Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J., 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol, 9(8), p.e1003118.

[80] Levy, J., Ravindran, A., Gross, D., Tamborindeguy, C. and Pierson, E., 2011. Translocation of 'Candidatus Liberibacter solanacearum', the zebra chip pathogen, in potato and tomato. Phytopathology, 101(11), pp.1285-1291.

[81] Li, L. and Steffens, J.C., 2002. Overexpression of polyphenol oxidase in transgenic tomato plants results in enhanced bacterial disease resistance. Planta, 215(2), pp.239-247.

[82] Licausi, F., Ohme-Takagi, M. and Perata, P., 2013. APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. New Phytologist, 199(3), pp.639-649.

[83] Liu, D., Johnson, L., and Trumble, J. T., 2006. Differential responses to feeding by the tomato/potato psyllid between two tomato cultivars and their implications in establishment of injury levels and potential of damaged plant recovery. Insect Science, 13(3), pp.195-204.

[84] Love, M.I., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), p.550.

[85] Major, I.T. and Constabel, C.P., 2008. Functional analysis of the Kunitz trypsin inhibitor family in poplar reveals biochemical diversity and multiplicity in defense against herbivores. Plant Physiology, 146(3), pp.888-903.

[86] McElrone, A.J. and Forseth, I.N., 2004. Photosynthetic responses of a temperate liana to Xylella fastidiosa infection and water stress. Journal of Phytopathology, 152(1), pp.9-20.

[87] Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W., 2003. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. The Plant Cell, 15(4), pp.809-834.

[88] Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y. and Herbst, E.V., 2005. The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. Plant physiology, 138(3), pp.1310-1317.

[89] Munyaneza, J.E., 2012. Zebra chip disease of potato: biology, epidemiology, and management. American Journal of Potato Research, 89(5), pp.329-350.

[90] Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J.B., Engel, A. and Fujiyoshi, Y., 2000. Structural determinants of water permeation through aquaporin-1. Nature, 407(6804), pp.599-605.

[91] Nachappa, P., Levy, J., Pierson, E. and Tamborindeguy, C., 2011. Diversity of endosymbionts in the potato psyllid, Bactericera cockerelli (Hemiptera: Triozidae), vector of zebra chip disease of potato. Current microbiology, 62(5), pp.1510-1520.

[92] Nachappa, P., Levy, J., Pierson, E. and Tamborindeguy, C., 2014. Correlation between “Candidatus Liberibacter solanacearum" infection levels and fecundity in its psyllid vector. Journal of invertebrate pathology, 115, pp.55-61.

[93] Noutoshi, Y., Kuromori, T., Wada, T., Hirayama, T., Kamiya, A., Imura, Y., Yasuda, M., Nakashita, H., Shirasu, K. and Shinozaki, K., 2006. Loss of necrotic spotted lesions 1 associates with cell death and defense responses in Arabidopsis thaliana. Plant molecular biology, 62(1), pp.29-42.

[94] Oerke, E.C., 2006. Crop losses to pests. The Journal of Agricultural Science, 144(01), pp.31-43.

[95] Olson, A.J., Pataky, J.K., Dárcy, C.J. and Ford, R.E., 1990. Effects of drought stress and infection by maize dwarf mosaic virus on sweet corn. Plant disease, 74(2), pp.147-151.

[96] Orellana, S., Yanez, M., Espinoza, A., Verdugo, I., Gonzalez, E., Ruiz-Lara, Simón. and Casaretto, J.A., 2010. The transcription factor SlAREB1 confers drought, salt stress tolerance and regulates biotic and abiotic stress-related genes in tomato. Plant, cell & environment, 33(12), pp.2191-2208.

[97] Pandey, P., Ramegowda, V. and Senthil-Kumar, M., 2015. Shared and unique responses of plants to multiple individual stresses and stress combinations: physiological and molecular mechanisms. Frontiers in plant science, 6, p.723.

[98] Pedley, K.F. and Martin, G.B., 2004. Identification of MAPKs and their possible MAPK kinase activators involved in the Pto-mediated defense response of tomato. Journal of Biological Chemistry, 279(47), pp.49229-49235.

[99] Ramegowda, V. and Senthil-Kumar, M., 2015. The interactive effects of simultaneous biotic and abiotic stresses on plants: mechanistic understanding from drought and pathogen combination. Journal of plant physiology, 176, pp.47-54.

[100] Ramegowda, V., Senthil-Kumar, M., Ishiga, Y., Kaundal, A., Udayakumar, M. and Mysore, K.S., 2013. Drought stress acclimation imparts tolerance to Sclerotinia sclerotiorum and Pseudomonas syringae in Nicotiana benthamiana. International journal of molecular sciences, 14(5), pp.9497-9513.

[101] Rejeb, I.B., Pastor, V. and Mauch-Mani, B., 2014. Plant responses to simultaneous biotic and abiotic stress: molecular mechanisms. Plants, 3(4), pp.458-475.

[102] Reymond, P., Weber, H., Damond, M. and Farmer, E.E., 2000. Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis. The Plant Cell, 12(5), pp.707-719.

[103] Rodrigues, A., Santiago, J., Rubio, S., Saez, A., Osmont, K.S., Gadea, J., Hardtke, C.S. and Rodriguez, P.L., 2009. The short-rooted phenotype of the brevis radix mutant partly reflects root abscisic acid hypersensitivity. Plant physiology, 149(4), pp.1917-1928.

[104] Romeis, T., Ludwig, A.A., Martin, R. and Jones, J.D., 2001. Calcium-dependent protein kinases play an essential role in a plant defence response. The EMBO journal, 20(20), pp.5556-5567.

[105] Roux, M., Schwessinger, B., Albrecht, C., Chinchilla, D., Jones, A., Holton, N., Malinovsky, F.G., Tör, M., de Vries, S. and Zipfel, C., 2011. The Arabidopsis leucine-rich repeat receptorâĂŞlike kinases BAK1/SERK3 and BKK1/SERK4 are required for innate immunity to hemibiotrophic and biotrophic pathogens. The Plant Cell, 23(6), pp.2440-2455.

[106] Shen, G., Pang, Y., Wu, W., Deng, Z., Zhao, L., Cao, Y., Sun, X. and Tang, K., 2006. Cloning and characterization of a flavanone 3-hydroxylase gene from Ginkgo biloba. Bioscience reports, 26(1), pp.19-29.

[107] Shinozaki, K. and Yamaguchi-Shinozaki, K., 1997. Gene expression and signal transduction in water-stress response. Plant physiology, 115(2), p.327.

[108] Shiu, Shin-Han, and Anthony B. Bleecker. Plant receptor-like kinase gene family: diversity, function, and signaling.Sci stke 113, no. re22 (2001): 1-13.

[109] Sreenivasulu, N., Sopory, S.K. and Kishor, P.K., 2007. Deciphering the regulatory mechanisms of abiotic stress tolerance in plants by genomic approaches. Gene, 388(1), pp.1-13.

[110] Suzuki, N., Rivero, R.M., Shulaev, V., Blumwald, E. and Mittler, R., 2014. Abiotic and biotic stress combinations. New Phytologist, 203(1), pp.32-43.

[111] Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., Yamaguchi-Shinozaki, K. and Shinozaki, K., 2002. Important roles of drought and cold-inducible genes for galactinol synthase in stress tolerance in Arabidopsis thaliana. The Plant Journal, 29(4), pp.417-426.

[112] Tanaka, H., Onouchi, H., Kondo, M., Hara-Nishimura, I., Nishimura, M., Machida, C. and Machida, Y., 2001. A subtilisin-like serine protease is required for epidermal surface formation in Arabidopsis embryos and juvenile plants. Development, 128(23), pp.4681-4689.

[113] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A., 2011. Differential expression in RNA-seq: a matter of depth. Genome research, 21(12), pp.2213-2223.

[114] Team, R.C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

[115] Thomma, B.P., Penninckx, I.A., Cammue, B.P. and Broekaert, W.F., 2001. The complexity of disease signaling in Arabidopsis. Current opinion in immunology, 13(1), pp.63-68.

[116] Vorwerk, S., Somerville, S. and Somerville, C., 2004. The role of plant cell wall polysaccharide composition in disease resistance. Trends in plant science, 9(4), pp.203-209.

[117] Wang, Z., Gerstein, M. and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), pp.57-63.

[118] Wu, C.T. and Bradford, K.J., 2003. Class I chitinase and β-1, 3-glucanase are differentially regulated by wounding, methyl jasmonate, ethylene, and gibberellin in tomato seeds and leaves. Plant Physiology, 133(1), pp.263-273.

[119] Xu, Z. and Zhou, G., 2008. Responses of leaf stomatal density to water status and its relationship with photosynthesis in a grass. Journal of experimental botany, 59(12), pp.3317-3325.

[120] Yamaguchi-Shinozaki, K. and Shinozaki, K., 1994. A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. The Plant Cell, 6(2), pp.251-264.

[121] Zhang, C.Q., Xu, Y., Lu, Y., Yu, H.X., Gu, M.H. and Liu, Q.Q., 2011. The WRKY transcription factor OsWRKY78 regulates stem elongation and seed development in rice. Planta, 234(3), pp.541-554.

[122] Zhang, X., Zou, Z., Gong, P., Zhang, J., Ziaf, K., Li, H., Xiao, F. and Ye, Z., 2011. Over-expression of microRNA169 confers enhanced drought tolerance to tomato. Biotechnology letters, 33(2), pp.403-409.

[123] Zhang, Z., Liu, X., Wang, X., Zhou, M., Zhou, X., Ye, X. and Wei, X., 2012. An R2R3 MYB transcription factor in wheat, TaPIMP1, mediates host resistance to Bipolaris sorokiniana and drought stresses through regulation of defense-and stress-related genes. New Phytologist, 196(4), pp.1155-1170.

[124] Giovannucci, E., Harlan, D.M., Archer, M.C., Bergenstal, R.M., Gapstur, S.M., Habel, L.A., Pollak, M., Regensteiner, J.G. and Yee, D., 2010. Diabetes and cancer: a consensus report. CA: a cancer journal for clinicians, 60(4), pp.207-221.

[125] Bao, B., Wang, Z., Ali, S., Ahmad, A., Azmi, A.S., Sarkar, S.H., Banerjee, S., Kong, D., Li, Y., Thakur, S. and Sarkar, F.H., 2012. Metformin inhibits cell proliferation, migration and invasion by attenuating CSC function mediated by deregulating miRNAs in pancreatic cancer cells. Cancer prevention research, 5(3), pp.355-364.

[126] Kisfalvi, K., Eibl, G., Sinnett-Smith, J. and Rozengurt, E., 2009. Metformin disrupts crosstalk between G protein-coupled receptor and insulin receptor signaling systems and inhibits pancreatic cancer growth. Cancer research, 69(16), pp.6539-6545..

[127] Lee, J.H., Kim, T.I., Jeon, S.M., Hong, S.P., Cheon, J.H. and Kim, W.H., 2012. The effects of metformin on the survival of colorectal cancer patients with diabetes mellitus. International Journal of Cancer, 131(3), pp.752-759.

[128] Tseng, C.H., 2012. Diabetes, metformin use, and colon cancer: a population-based cohort study in Taiwan. European Journal of Endocrinology, 167(3), pp.409-416.

[129] Zhou, G., Myers, R., Li, Y., Chen, Y., Shen, X., Fenyk-Melody, J., Wu, M., Ventre, J., Doebber, T., Fujii, N. and Musi, N., 2001. Role of AMP-activated protein kinase in mechanism of metformin action. The Journal of clinical investigation, 108(8), pp.1167-1174.

[130] Zakikhani, M., Dowling, R., Fantus, I.G., Sonenberg, N. and Pollak, M., 2006. Metformin is an AMP kinase-dependent growth inhibitor for breast cancer cells. Cancer research, 66(21), pp.10269-10273.

[131] Dowling, R.J., Zakikhani, M., Fantus, I.G., Pollak, M. and Sonenberg, N., 2007. Metformin inhibits mammalian target of rapamycin-dependent translation initiation in breast cancer cells. Cancer research, 67(22), pp.10804-10812.

[132] Hua, J., Sima, C., Cypert, M., Gooden, G.C., Shack, S., Alla, L., Smith, E.A., Trent, J.M., Dougherty, E.R. and Bittner, M.L., 2012. Tracking transcriptional activities with high-content epifluorescent imaging. Journal of biomedical optics, 17(4), pp.0460081-04600815.

[133] Team, R.C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

[134] Hirsch, H.A., Iliopoulos, D., Tsichlis, P.N. and Struhl, K., 2009. Metformin selectively targets cancer stem cells, and acts together with chemotherapy to block tumor growth and prolong remission. Cancer research, 69(19), pp.7507-7511.

[135] Erices, R., Bravo, M.L., Gonzalez, P., Oliva, B., Racordon, D., Garrido, M., Ibañez, C., Kato, S., Brañes, J., Pizarro, J. and Barriga, M.I., 2013. Metformin, at concentrations corresponding to the treatment of diabetes, potentiates the cytotoxic effects of carboplatin in cultures of ovarian cancer cells. Reproductive Sciences, 20(12), pp.1433-1446.

[136] Krech, T., Thiede, M., Hilgenberg, E., Schäfer, R. and Jürchott, K., 2010. Characterization of AKT independent effects of the synthetic AKT inhibitors SH-5 and SH-6 using an integrated approach combining transcriptomic profiling and signaling pathway perturbations. BMC cancer, 10(1), p.287.

[137] Venkat, P. S., Krishna R. N., and Aniruddha D., 2017. A Bayesian Network-Based Approach to Selection of Intervention Points in the Mitogen-Activated Protein Kinase Plant Defense Response Pathway. Journal of Computational Biology 24(4), pp.327-339.

[138] Venkatasubramani, P., Sima, C., Hua, J., Cypert, M., Bittner, M. and Datta, A., 2017. Epifluorescent imaging study of the effect of anti-diabetic drug metformin on colorectal cancer cell lines in vitro. J Cancer Res Ther, 5(4), pp.19-23.